communications biology

Perspective

https://doi.org/10.1038/s42003-024-07212-3

Improving the reporting of metagenomic virome-scale data

Wei-Shan Chang [1,2], Erin Harvey [1], Jackie E. Mahar [1,3], Cadhla Firth [4], Mang Shi [5], Etienne Simon-Loriere [6], Jemma L. Geoghegan [7,8] & Michelle Wille [1,9] ✉

Over the last decade metagenomic sequencing has facilitated an increasing number of virome-scale studies, leading to an exponential expansion in understanding of virus diversity. This is partially driven by the decreasing costs of metagenomic sequencing, improvements in computational tools for revealing novel viruses, and an increased understanding of the key role that viruses play in human and animal health. A central concern associated with this remarkable increase in the number of virome-scale studies is the lack of broadly accepted "gold standards" for reporting the data and results generated. This is of particular importance for animal virome studies as there are a multitude of nuanced approaches for both data presentation and analysis, all of which impact the resulting outcomes. As such, the results of published studies can be difficult to contextualise and may be of reduced utility due to reporting deficiencies. Herein, we aim to address these reporting issues by outlining recommendations for the presentation of virome data, encouraging a transparent communication of findings that can be interpreted in evolutionary and ecological contexts.

The rapid expansion of metagenomic studies has led to a revolution in virology

Metagenomics has revolutionised the field of virology, allowing the rapid detection and genomic characterization of known and novel viruses from diverse environments. The metagenomic revolution has revealed that viruses are likely the most abundant biological entity on the planet and viral diversity extends beyond that predicted prior to the genomic era[1]. As well as virus discovery, metagenomic sequencing has substantially expanded our understanding of the host range of virus families. For example, the *Orthomyxoviridae*[2–5] and *Flaviviridae*[6,7], which were classically defined as mammalian-infecting viral families, are now known to infect a wide range of hosts, including diverse invertebrate phyla.

With decreasing sequencing costs, increasing power of computational resources, and the expansion and development of bioinformatic tools over the last 20 years, there has been a corresponding increase in the number of virome characterisation and virus discovery studies using metagenomics (Fig. 1). Here, we refer to metagenomics as the high-throughput sequencing (HTS) of the total genetic material within a sample, in which metatranscriptomics is the specific HTS of RNA. This technique has led to the popularisation of the term 'virome' to refer to the total diversity of viruses present in a given sample. Indeed, the number of virome papers published per year has increased from 44 in 2013 to 388 in 2023 (Fig. 1) and continues to rise. Associated with this increase are the approximately 750,000 uncultivated viral genomes identified in metagenomic data sets between 2016-2018[8] and a 7-fold increase in the number of novel virus sequences added to GenBank in the decade following the launch of the Illumina HiSeq platform, from 2010 (n = 1,053) to 2020 (n = 7,016) (Fig. 1). As the cost of sequencing continues to decrease, these numbers will likely continue to rise apace in the coming years.

However, compared to studies of the microbiome, or bacterial communities, the integration of metagenomics into virology research is in its infancy. Microbiome research was transformed by amplicon sequencing of the highly conserved 16S ribosomal subunit gene found in all bacteria; not only did this lead to important research findings, but it drove innovation in development of tools and technology to facilitate microbiome studies beyond 16S into whole genome sequencing[9–11]. While arguably still the gold

[1]School of Medical Sciences, The University of Sydney, Sydney, NSW, Australia. [2]Health and Biosecurity, Commonwealth Scientific and Industrial Research Organisation, Canberra, ACT, Australia. [3]Australian Animal Health Laboratory and Health and Biosecurity, Commonwealth Scientific and Industrial Research Organisation, Geelong, VIC, Australia. [4]College of Public Health, Medical, and Veterinary Sciences, James Cook University, Townsville, Australia. [5]Sun Yat-Sen University, Shenzhen campus of Sun Yat-Sen University, Shenzhen, China. [6]Evolutionary Genomics of RNA Viruses, Institut Pasteur, Université Paris Cité, Paris, France. [7]Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand. [8]Institute of Environmental Science and Research, Wellington, New Zealand. [9]Centre for Pathogen Genomics, Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, VIC, Australia. ✉e-mail: michelle.wille@unimelb.edu.au

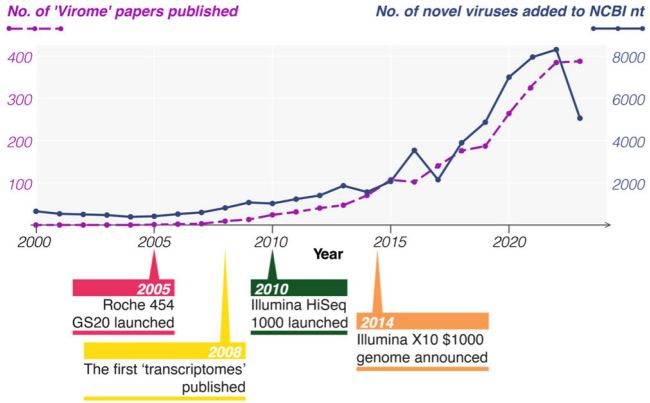


Fig. 1 | Rapid expansion of metagenomic-based virome studies and novel viral sequences over time. In violet: The number of studies published in NCBI's PubMed database each year from 2000 to 2023 that report metagenomic virus discovery/virome analyses [Search query: (metagenomic OR metatranscriptomic) AND (virus OR virome)]. In blue: The number of new virus organisms published in NCBI's nucleotide database each year from 1 January 2000 to 31 December 2023, sorted by species name. Below the graph, key events in the development of metagenomics are indicated.

standard, the reliance on traditional culture or microscopy methods severely limits our capacity to study the true diversity and abundance of viruses. As virome scale research is more widely undertaken, and standardized protocols and data analysis structures are developed, the field is on the same trajectory as microbiome research. Indeed, metagenomics is a cornerstone of research in microbiology today[12].

Current pitfalls and challenges of metagenomics

The rapid growth of viral metagenomics has been accompanied by a similar expansion of tools and techniques for data analysis and reporting, with no clear consensus on best practices. The lack of a standardised approach is unsurprising given that metagenomic studies consider hugely different host taxa and commonly pose very different research questions. This is further complicated by the all ever-increasing complexity of taxonomic assignments. In addition, new tools and approaches are continuously developed to handle the unique challenges of working with virome scale data, such as lack of appropriate databases, few tools for mining segmented viruses, and no standards for sequence clustering, which have been highlighted in a recent consensus statement[13]. Indeed, more than 15 new pipelines and packages for virus discovery have been reported in 2023[14–29], excluding custom approaches. In addition to tools, stand-alone databases including sequence data, functional annotations and metadata are emerging, which can be incorporated into the diversity of workflows[30]. While this diversification is expected in a growing field and will lead to methodological improvements, the lack of standardized approaches and an absence of appropriately detailed reporting limits the ability to compare and replicate studies, potentially decreasing their value to the scientific community. To address these deficiencies a more systematic approach to data collection, reporting, and analysis is clearly required in the field of viral metagenomics.

Current studies in virus metagenomics can be limited in a number of ways. Methods sections can lack the detail required for reproducibility, contextualization and evaluation. A variety of approaches to sample preparation and data analysis may be adopted depending on sample type and the specific aims of the project, invariably impacting the study outcome. For example, the use and type of viral enrichment or host depletion techniques (e.g. particle filtration, nuclease digestion, rRNA depletion) varies widely, or may not be performed at all. Different extraction kits will similarly alter the detectability and abundance of different types of viruses depending on the methods used[31,32]. Another common laboratory practice is pooling multiple samples prior to sequencing, yet the pooling strategy is sometimes described in insufficient detail to be repeated. This also applies to bioinformatic workflows for sequence analysis. For example, some pipelines assemble only sequences (*i.e.* 'reads') that have been identified as viral through sequence similarity searches like BLAST, while others will assemble all reads prior to sequence identification. This choice impacts the assembly of highly divergent viruses, biasing downstream estimates of viral diversity, community composition, etc., that should ideally be comparable across studies. The methodological approach to estimating viral abundance from read counts will have similarly important effects on downstream ecological analyses. While it may be theoretically possible to account for some variation in data collection and analytical approaches when performing cross-study comparisons, this is currently impractical due to a lack of detailed methodological reporting.

Insufficient detail in the sharing of metagenomic data, associated metadata, and the reporting of analytical results is also commonplace. For example, despite existing checklists[33], accompanying metadata (e.g., collection date, location, host, sample type, disease state) may be excluded or not comprehensive[13], limiting the ability to place them in the correct ecological or evolutionary context. The use of raw data outputs from automated bioinformatics pipelines has led virus discovery studies that provide little information on the virus beyond its unannotated genomic sequence. This can be resolved through sequence annotation and the inclusion of phylogenies as discussed below.

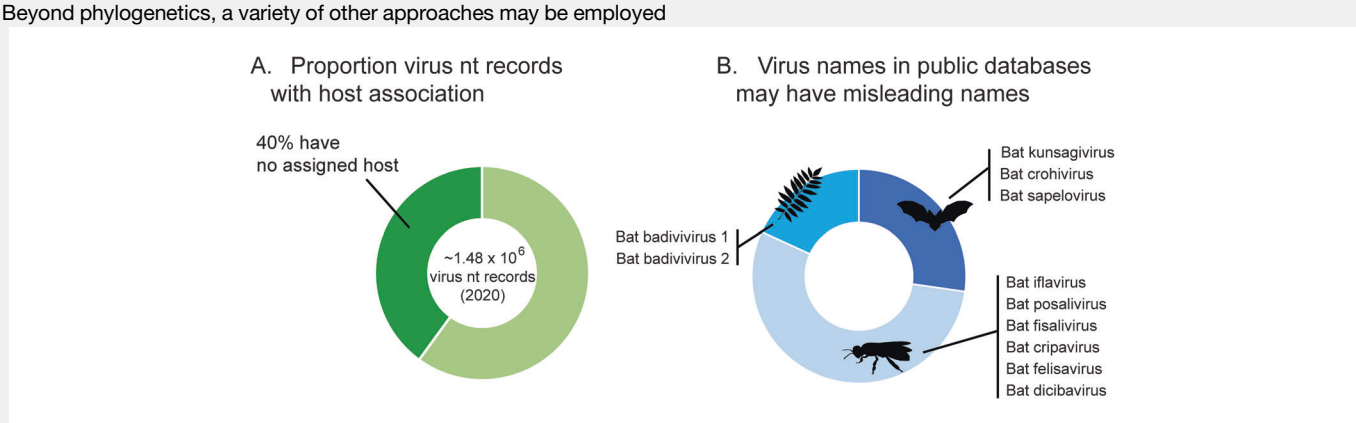
A key challenge in viral metagenomics is the correct association of a viral sequence with its host; however, host-virus associations are often challenging, sometimes neglected altogether, or incorrectly reported. This may be because the host of a particular virus is not necessarily the species from which it was sampled, as many viruses in a metagenomic sample originate from the sampled species' microbiome, or diet, or simply result from laboratory contamination[34]. This is described in more detail in Box 1. Determining host associations is made more complex depending on how viral sequences are named. In particular, the name of the sampled organism is often included in novel virus names which can be misleading when the sampled species has not been determined as the definitive host, as recognised by the ICTV[35,36] (Box 2). For example, neither Bat Iflavirus (GenBank Accession NC_033823) nor Goose Dicistrovirus (GenBank Accession NC_029052) have reservoirs in vertebrates. Rather, these viruses are likely associated with the invertebrates comprising the diet of the sampled vertebrate hosts. Erroneous host associations in public databases can lead to cascades of host mischaracterization, and have the potential to result in incorrect evolutionary or ecological inferences.

Variation in the approaches used for metagenomic data analysis is equally problematic for the interpretation of virome data, particularly when limited methodological detail is provided. As a large proportion of the virosphere remains unresolved[37], virome characterisation is often complex and requires careful analysis. For example, many virome studies report viruses without conducting phylogenetic analyses, although this is central to virus classification and the baseline for many evolutionary and ecological inferences (e.g.[38,39]). Providing viral gene sequences can be reliably aligned, phylogenetics is arguably the best way to validate novel viral sequences and determine their taxonomy, while also providing information on the likely host or whether the virus may be a contaminant[34] (see Box 3). Yet this step is sometimes omitted, and genetic characterisation conducted using only broad-scale summary statistics and similarity-based analyses (e.g. BLAST) that average over a large number of parameters and which do not result in analytical precision. Examples include a reliance on diversity metrics such as pi, richness, Shannon diversity index and/or characterising viral operational taxonomic units (vOTU), or the identification of sequence clusters though sequence similarity alone, which are problematic when performed without contig verification such as virus species identification or sequence annotation[38,40–44]. The consequence of presenting only diversity metrics and not performing genome annotation is that the viruses in question may not be deposited into sequence repositories like GenBank, or are deposited with no annotation, no taxonomy information and uninformative names such as 'unclassified Riboviria'. Over time, this reduces the utility of the public databases that form the basis of novel virus identification[45–47]. As the

BOX 1 | Host associations

Clarifying potential host associations is of critical importance to revealing the virosphere, and at a lower level, identifying relationships in the context of "host-pathogen" networks. Inaccuracies in host association and/or naming viruses after hosts despite incongruous host-association (Box Fig. 1) lead to numerous problems for not only the study in question, but for the community who rely upon these data for taxonomy, or ecological and evolutionary questions.

Identifying host associations is often challenging, but progress is being made through machine learning approaches for viruses which infect bacteria and archaea (e.g.[70]). Less progress has been made for viruses infecting animals, and a number of different approaches have been put forward, as summarised in Cobbin et al. [34]. The most straightforward is to conduct phylogenetic analysis to identify the host association of closely related viruses, assuming virus-host codivergence (being mindful that hosts could be misassigned in the database)[71]. Beyond phylogenetics, a variety of other approaches may be employed including: (1) exploring signatures of virus and host genome coevolution by comparing the virus and potential host codon pairing and/or oligonucleotide frequency patterns[72]. (2) Correlating viral abundance with the abundance of intra-host microbe marker genes in cases where metatranscriptomic sequencing has been employed. (3) Conducting large-scale virus-host association studies in which diverse host data sets from resources such as the Sequence Read Archive (SRA) are mined for viruses, and (4) excluding or identifying the presence of other potential host species by analysing non-viral sequences from the same library. Tools using a combination of approaches are showing utility and high levels of accuracy. For example, a machine learning model using a combination of phylogenetics and biases in viral genome composition was successfully used to identify arthropod vectors for a substantial array of viruses[73].



Box Fig. 1 Current state of host association. A As indicated in dark green, a substantial proportion of viral records in GenBank do not have an associated host. Modified from Cobbin et al. [34]. **B** Eleven picornaviruses recovered from bat faeces, all including "bat" in the virus name, have 3 different hosts as indicated by silhouettes overlaying the plot. Modified from Yinda et al. [69].

proportion of metagenomic data in these databases continues to increase, it will be vital that sequences are properly characterised, and that this characterisation is clearly reported.

Taken together, a large diversity of tools and approaches are used in studies underpinned by virome-scale metagenomic data, and gaps in reporting results and methodologies may limit the value of time-intensive and costly metagenomic studies to the scientific community. The deposition of poorly characterised sequence data into public databases, may even detrimentally impact subsequent studies. A consensus on how to report virome-scale metagenomic data is clearly warranted.

Current standards for the presentation of metagenomic studies and their short comings

Without specific guidelines, most genome sequences in databases are sparsely annotated with the information required to guide data interpretation and knowledge generation[33]. As a result, an array of checklists comprising minimum standards for sequence-associated metadata reporting have been outlined and made available by the Genomics Standards Consortium (https://www.gensc.org/pages/standards/checklists.html). An abbreviated summary of the checklist relevant to the data produced in virome-scale studies is presented in Fig. 2 and includes the Minimum information about a marker gene sequence (MIMARKS) checklist as an extension to the Minimum Information about any Sequence (MIxS) list[33], the Minimum Information to report Uncultured Virus Genome (MIUViG)[8], and recommendations presented in Ladner et al. [48]. These checklists provide a useful starting point for developing a comprehensive set of recommendations for the presentation of virome-scale data analysis and the resulting genomes.

Briefly, MIxS encompasses genome and metagenome sequences, marker gene sequences, and single-amplified and metagenome-assembled bacterial and archaeal genomes. This checklist is borne out of the Minimum Information about a Genome Sequence (MIGS) and Minimum Information about a Metagenome Sequence (MIMS) and includes metadata and technology specific checklists[33,49]. A useful extension of the MIxS checklist is the MIMARKS checklist[33]. Together, these checklists suggest the inclusion of metadata regarding the following: (1) Data and investigations – data submission to public database(s) and basic description of the project name, (2) Environment information – collection date, geographic location, features and materials, (3) Nucleic acid sequence source, which refers to the general sequencing approach and is useful as a common standard to convey the quality, and therefore utility, of the associated genome sequences, and (4) Sequencing platform, technology, and basic bioinformatic tools, such as those relevant for assembly (Fig. 2).

Current minimum standards recommendations for metagenome assembled or uncultured genomes include the MIMAG (Minimum Information about a Metagenome-Assembled Genome sequence) (Bowers et al. and MIUViG[8]. The former is targeted specifically toward bacterial genomes, while the MIUViG checklist, particularly when combined with the recommendations of Ladner et al. [48], are more oriented to viral data sets. Together, they provide suggestions for inclusion of the data source and quality, software for analysis of assembly, virus identification, annotation,

BOX 2 | Improving clarity in virus presentation

Virus taxonomy and species naming is under the purview of the International Committee on the Taxonomy of Viruses (ICTV; https://ictv.global/)[74], and virus organism names and virus taxon names are not necessarily the same[36]. Currently we are seeing a substantial and continued overhaul of virus taxonomy and nomenclature, with changes occurring on different timelines among the different subcommittees/virus families (*e.g.*[75]). This is creating substantial challenges in presenting both novel and established virus species in scientific articles. This is further causing confusion around virus names already present in databases. For example, "bat crohivirus" presented in Box Fig. 1, is not a member of the genus *Crohivirus*. Herein, we provide some suggestions to improve the clarity of virus names presented in studies, but with the caveat that it's a continually evolving landscape.

For clarity only, it is preferable to provide highly divergent virus sequences that potentially constitute new species with a unique virus name. It is important to note that virus names provided by the author are not synonymous with virus species names, which are decided by the ICTV following assessment and ratification of novel viruses[74]. Using contig names or complex coded names are not ideal as they may be impossible to decipher by others wishing to include sequences for comparison in future studies. Current guidelines indicate that viruses cannot be named after locations, host species, or copyright protected names[35], and only after people with certain caveats (https://ictv.global/about/code). Location and host species should not be included in virus names as the point of detection may not be a true reflection of spatial or host range. While using location in the name is still done for bacterial viruses, it is argued against for putative animal pathogens. It is important to confirm that proposed names have not been used prior. Presentation of previously described viruses should be done in accordance with both the ICTV and field-specific nomenclature. Virus species names and taxonomy should be formulated as outlined by Simmonds et al. [62], and presented as outlined by Zerbini et al. [36]. Providing additional detail regarding clades, variants, strains, subtypes or genotypes within established nomenclature systems is crucial for improving the value of the presented findings.

BOX 3 | Phylogenetic analysis as a key step in virus verification

Given most of the virosphere remains undiscovered[37], validation and characterisation of novel viral contigs is imperative, and relying on only viral operational taxonomic units (vOTU's), diversity statistics, or BLAST results is not sufficient. To validate novel viral sequences, reveal their relationship to other viruses, and better assign putative taxonomic ranking, a robust phylogenetic analysis is required. Indeed, a key challenge is the mismatch between results from bioinformatic tools relative to hierarchical taxonomic structure across all ranks, which may conflict with ICTV-ratified taxa that have been meticulously defined by experts, and therefore careful verification through phylogeny is required[76]. Pipelines relying only on BLAST have a severe shortcoming, particularly in the context of highly divergent virus sequences. These divergent virus sequences may have <40% amino acid similarity, and therefore the closest relative identified by BLAST is highly approximate. Indeed, the sequence that is listed first in a BLAST output is not necessarily the closest relative according to a phylogenetic analysis, and this has key ramifications for inferences utilizing, and for reporting of, taxonomy, particularly of divergent viruses. Unlike BLAST which can report results for matches based on only a short region of the sequence, phylogenetics is based on the complete length of the alignment provided, which usually includes the entire novel sequence or entire translated product of a conserved gene, like the RNA-dependent RNA polymerase (RdRp). Unlike BLAST, phylogenetic analysis is central to revealing lower-level classifications (*e.g.* genus, species, lineage level) and can provide insight into potential viral characteristics based on its closest relatives (*i.e.* likely host, potential for virulence, whether it may be a contaminant *etc.*). An example of taxonomic discrepancy between BLAST and phylogenetic results comes from the original description of Bruthen virus, an unassigned member of the *Bunyavirales*[77]. If BLAST based analysis were being utilized, this virus would be classified as a member of the *Phlebovirus* genus, with 25.4% amino acid similarity to a tick-borne zoonotic virus *Dabie bandavirus* (previously Huaiyangshan virus), and thus of biological relevance to the avian host. Phylogenetic analysis revealed this virus did not fall into the genus *Phlebovirus*, but was rather a divergent virus of the *Bunyavirales*.

Phylogenetic analysis is also central to ascertaining host associations (expanded upon in Box 1). In studies relying on sample types such as faecal samples, it can be challenging to ascertain true host-virus associations: viruses found in faecal samples could comprise viruses of the host, microbiome, or diet. As there is often long-term co-divergence between hosts and viruses, viral phylogenies can be highly structured by host taxonomy, and therefore, many host inferences can be made based upon phylogenetic placement. For example, within the genus *Flavivirus*, host and vector associations can be phylogenetically derived, such that it is possible to reveal whether viruses are likely vector-borne or arthropod specific[6,78]. Similarly, phylogenetic analysis can also be used to identify sequences from laboratory contamination, which appears to be commonplace and can comprise a wide variety of viruses[59,60]. Specifically, phylogenetic trees can reveal whether contigs are in clades dominated by confirmed lab contaminants or whether contigs are incorporated into clades associated with known hosts.

structure, completion of a high-quality draft virus genome, contaminating agents, etc.

Although they provide an important foundation, these checklists lack recommendations on study aims or specific downstream analyses of viral contigs that include phylogenetic verification or ascertaining host associations (Fig. 2), which we will address below. Overall, the current minimum standards checklists and recommendations are not sufficiently comprehensive when applied to virome-scale data. We therefore propose an increase in scope to the existing checklists, and provide suggestions on how specific recommendations may be implemented into virus discovery, evolution, and ecology studies (Fig. 3, Supplementary File 1).

10 recommendations for reporting virome-scale studies

1. Sample collection, storage, transport, and metadata

Information is required on materials used for the collection and storage of samples (e.g., type of swab, transport media, etc.), as each can have important consequences for nucleic acid quality[50]. Sample metadata should

	MIMARKS/MIxS	MIUVIG	Standard virus genome
Project investigation	Description of investigation type and Project name		
Data repository	Submitted to INSDC (SRA, DRA, GenBank, ENA, DDBJ.)		Repositories of reads and genomic information (i.e. GenBank)
Sample collection metadata Environment and source descriptors	Collection date Geographical location (latitude and longitude) Environmental biome Environmental features Environmental materials Host association with applicable environmental package (animal associated, human associated, sediment, soil, wastewater)	Source of UViG (type of dataset)	
Sequencing technology and locus	Target gene sequencing (16s, 18s rRNA) or locus name for marker gene Sequencing methods (Sanger, Illumina, etc)		
Genome assembly		Tools/Software used for assembly including version number, parameter and cut-offs Assembly quality and genome quality: (1) finished (2) high-quality draft genome (3) genome fragments with annotation Number of contigs	Assembly quality and genome quality: (1) complete with full genome (2) high-quality draft genome (3) coding complete with complete ORFs Number of contigs
Virus identification and genome characterization		Tools/Software used for virus identification including version number, parameter and cut-offs Prediction genome type and structure Virus operational taxonomic units (%ID)	
Contamination analysis		Contamination threshold suggested	Sequencing of blank control

Fig. 2 | Current minimum standards, and how they may be applied to metagenome-assembled viral genomes. Standards outlined in MIMARKS/MIxS are those outlined in Yilmaz et al. [33], those from MIUVIG are those outlined in Roux et al. [8], and those included in a standard virus genome are those outlined in Ladner et al. [48]. INSDC International Nucleotide Sequence Database Collaboration, SRA Sequence Read Archive, DDBJ DNA Data Bank of Japan, UViG Uncultivated Virus Genome, DRA DDBJ Sequence Read Archive, ENA European Nucleotide Archive, rRNA ribosomal RNA, ORF Open Reading Frame.

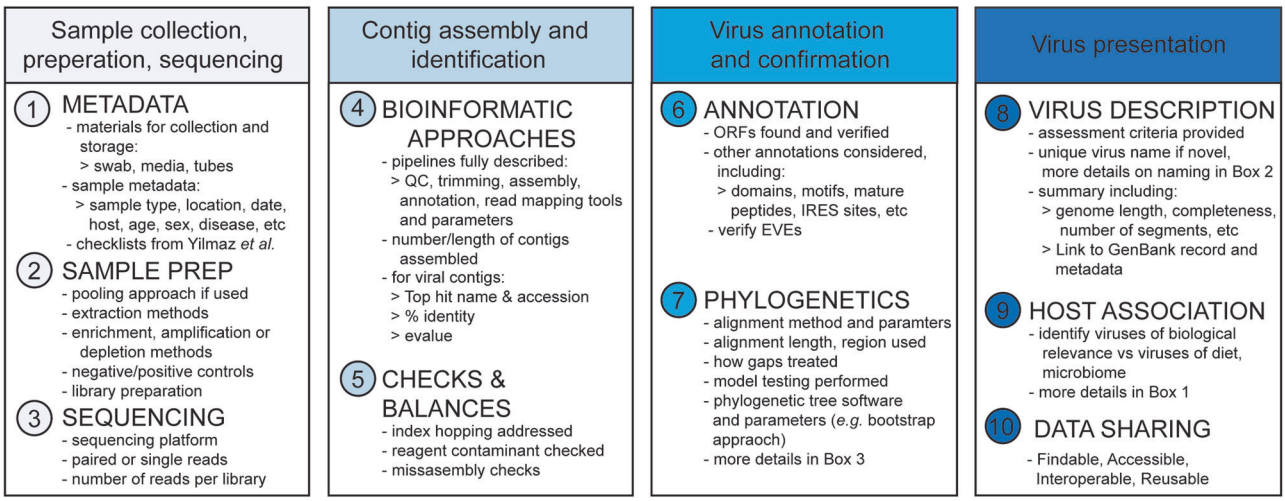


Fig. 3 | Summary of data presentation features we propose for inclusion in all virome studies. A tabular checklist is provided as Supplementary File 1.

include sample type, location and date of sampling, and sampled organism, as well as other biologically relevant data depending on the aim of the study and any ethical considerations. For example, age[51], sex[52], season[53], disease status[54], and phenotypic characteristics[55,56] all have the potential to influence the virome and may be relevant to a particular study. Detailed metadata checklists presented in Yilmaz et al. [33] should be used, and can be downloaded from https://www.gensc.org/pages/standards/checklists.html.

2. Sample preparation and viral enrichment or depletion protocols

Details of nucleic acid extraction methods, virus enrichment, amplification, or depletion protocols, sequence library preparation, and negative/positive controls should be presented. A description of the approach used for sample pooling should be presented if relevant. Sample preparation approaches, such as pooling, can affect the interpretation of results, such as calculations of viral sequence abundance and richness.

3. Sequencing methodology

A description of the sequencing methodology, including platform, read length, and whether paired- or single-end or stranded or non-stranded approaches have been performed provides important information. This should also include results on the number of sequence reads generated per library.

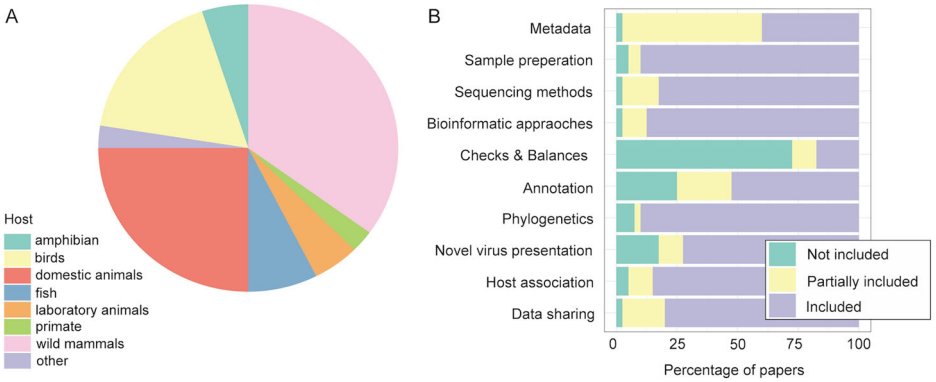


Fig. 4 | Papers published in 2023 using virome scale methods of non-human vertebrate hosts demonstrate many of our recommendations are already being considered by the community. A Pie chart of the hosts of virome studies assessed here. **B** Detailed assessment of the 10 recommendations proposed here in studies performed in animal hosts. The scoring system included whether each recommendation was (i) fully included as stated here, (ii) partially included such that only some aspects of the recommendation were incorporated, or (iii) whether the recommendation was not considered.

4. Bioinformatic approaches

Bioinformatic pipelines should be reproduceable where possible. All details around software, parameter settings and manual steps should be described. Details regarding quality control, trimming, assembly, contig annotation, and read mapping provide valuable information. The bioinformatic approaches used for taxonomic assignment of contigs (or reads) should be specified. For contigs of interest (*i.e.* those comprising viruses), the results of sequence similarity searches (e.g. BLAST) including closest genetic relative, percent sequence identities, alignment lengths, e-values, and contig lengths should also be provided in the main text.

5. Methodological checks and balances

"Index hopping" (wherein a proportion of reads are incorrectly indexed, usually 0.01-0.1% of reads if using common Illumina technologies) should be accounted for during data analysis (e.g.[57,58]). Efforts should be made to confirm that viral contigs were not derived from reagents or incidental contaminants. This can be achieved by comparing the results to lists of known reagent contaminants[59,60] as well as to experiment-specific no-template (negative) controls. Finally, steps to detect assembly errors, such as mapping reads back to viral contigs, and identification of appropriate functional domains, should be taken and reported. PCR confirmation may also be used as a verification step for metagenomic data, especially in cases where read mapping suggests a potential misassembly, where viral genome organisations diverge greatly from the structure expected, or where reagent contamination is suspected.

6. Annotation of viral transcripts

Open reading frames should be identified and verified as potential viral proteins based on conserved domains, signature motifs, and sequence homology with related viruses. If full viral genomes are identified, additional annotations may include the identification of prominent motifs and domains (such as the RdRp, helicase, and protease), mature peptides, and internal ribosomal entry sites, amongst others. In cases where segmented viruses are revealed, approaches as to how segments were assigned to viruses should be provided. Ideally, an attempt to identify and annotate endogenous viral elements (EVEs) should also be made and the approaches used reported, such as identifying truncated and/or non-functional proteins, investigating the genomic context from DNA sequencing, and using dedicated software (e.g.[61]).

7. Phylogenetic analysis of putative viral transcripts

Phylogenetic analysis of newly identified viral transcripts is the gold standard for virus classification and should include sequences at the appropriate taxonomic level required to classify a given virus. For example, if the virus is a new detection of an established species, relevant members of the virus species should be included. If the virus is divergent enough that it may constitute a new species, it is important to include other members of the genus, family, or order to provide adequate context (expanded upon in Box 3). The genomic region or protein used, alignment length, and tools used for sequence alignment should be reported, along with information on the methods used for the removal of poorly aligned regions, model testing, phylogenetic inference, and nodal support estimates (e.g. bootstrapping).

8. Presenting putatively novel viruses

When considering assigning taxonomy to newly characterized viruses, the thresholds of nucleotide and/or amino acid similarity used for classification should be reported. The ICTV criteria for the demarcation of viral species are usually defined by varying percent nucleotide or protein similarity thresholds depending on the viral families and genera, may be based on different genes/proteins (or complete genomes), and may incorporate other (non-sequence) information. It is important to note that as the ICTV officially designates species and associated species names (Box 2), any virus names proposed by the study authors constitute the sequence or virus organism name. In addition to sequence name, putative taxonomy should be included[62]. The presentation of new viruses should also include data on contig lengths, genome coverage and completeness, the number of segments recovered, and a link to the GenBank record and associated metadata. In the case where transcriptomic data were used, methods used to calculate viral abundance should be presented.

9. Virus-host associations

True virus-host associations are often difficult to determine and need to be carefully considered, particularly in the context of sample type (*i.e.* tissue versus faeces). For example, gut and cloacal samples are likely to include viruses that are biologically relevant for the host, as well as viruses associated with diet, the environment, and the microbiome. Viruses should be presented in the context of host association to avoid cases in which a disease is incorrectly attributed to a novel virus detection that is not biologically relevant. At a minimum, phylogenetic analysis (point 7) should be used to assess the potential host, but additional methods that can be utilised to determine the likely host are discussed in detail in Cobbin et al. [34], and in Box 1.

10. Data sharing principles: Findable, Accessible, Interoperable, Reusable

Sequencing reads should be made available on the Sequence Read Archive (SRA) or an equivalent open access database, with consideration to data sovereignty if applicable. Assembled viral genome sequences should be published in an International Nucleotide Sequence Database Collaboration (INSDC) database (such as GenBank or ENA). Ideally, sequences will be deposited alongside taxonomy, metadata and with appropriate annotations to improve utility, and must be linked to the deposited sequencing reads, as outlined in Adriaenssens et al. [47]. ORF translations should also be included in the GenBank/ENA record for the sequence to be included in NCBI protein database. It is also good practice to ensure that newly developed bioinformatic approaches or pipelines used for data analysis are made freely available on open-source platforms such as GitHub (or upon request). It would also be beneficial to the research community to upload laboratory protocols or workflows to repositories that provide persistent identifiers (e.g.

DOIs) such as protocols.io. Unique identifiers for each data set, metadata set, or manuscript should be clearly linked, readily findable and available for use to ensure alignment to Open Data Science Goals.

Community use of proposed guidelines

We have provided a potential road map for metagenomic virome-scale data reporting through recommendations that build on components already reported in a substantial proportion of studies, with key foundations in available minimum standards checklists. Importantly, the road map provided here can accommodate the diversity of laboratory and bioinformatic approaches currently employed in virome research, yet is flexible enough to accommodate future innovations in the field.

To assess current community practices, we examined all virome-scale studies published in 2023, focussing on vertebrate animal systems. Specifically, we identified studies focussing on non-human vertebrates (n = 40) in all PubMed hits for "virome" (n = 471). Overall, we found that most studies included details of sample preparation, sequencing methods and bioinformatic approaches, either in detail or partially (Fig. 4). Across our 10 recommendations, we found the lowest uptake was on "checks and balances" (recommendation 5), which comprises the inclusion of no template control libraries to identify putative reagent contamination, and addresses index hopping. However, as the field of viral metagenomics matures, so too will our appreciation of the limitations of the associated tools and techniques, and as a result, more and improved checks and balances will be incorporated. As such, the low uptake of this recommendation is most likely a reflection of an area where there is the largest capacity for improvement. Also of note was that more than a quarter of studies failed to include information on virus annotation (Fig. 4). Notably, only 5 studies included all items fully[63–67], demonstrating the need for ongoing improvements.

The current inconsistency in methods and results reporting is most likely the direct result of a lack of recommendations available in this rapidly expanding field. We anticipate that the unified and inclusive framework we have presented here will be substantially more straightforward and accessible (i.e., if you build it, they will come), and in turn will lead to a substantial improvement in the utility of virome-scale metagenomic research. The five papers reviewed here that incorporated all 10 of our recommendations[63–67], may serve as useful examples that demonstrate the appropriate application of the proposed standards. For example, all studies include comprehensive metadata, including species, locations, disease status, age (when known), swab and media types, etc. Brito et al. [66] sequenced not only diseased, but also healthy controls to put results into better context. Costa et al. [64] clearly outlines how putative false positives were addressed through additional searching of translated ORFs, contaminants were ruled out using Check V, and to confirm that no missassembly occurred, reads were mapped back with bowtie 2. Costa et al. further used RT-PCR to validate vertebrated associated viruses, providing substantial confidence in the results generated. Wierenga et al. [63,67] presented clear annotation, phylogenetic analysis and novel virus presentation. All papers undertake host association, although the approaches vary. Overall, through comprehensive reporting, the results of these studies are accessible.

Conclusion

There is a lack of consensus on how best to perform virome-scale metagenomic research, a problem exacerbated by a lack of sufficient methodological detail in some publications. We have provided a set of possible guidelines for the presentation of virome-scale data that will provide a foundation for better practices in data analysis and presentation, improving the usefulness of the results for the scientific community. As virome-scale studies are relatively new, we expect that new methods and approaches to data generation and analysis will continue to be developed. However, without a solid foundation of unifying guidelines underlying a set of best practices, these studies cannot be compared or sufficiently evaluated. For example, in 2009 following the explosion of quantitative PCR (qPCR) as a tool for everything from disease surveillance to gene expression studies, a comprehensive set of guidelines were produced (the MIQE guidelines) which have had a positive and overarching impact on all studies using qPCR[68]. We believe that the guidelines provided here are timely and will provide a clear benefit by unifying best-practices on virome-scale studies and alleviating current shortcomings in the presentation of results, while also providing a useful resource for newcomers to the field.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data needed to replicate the analysis and findings of this study are available here: https://doi.org/10.5281/zenodo.13271258.

Code availability

Code for Fig. 4 available at https://github.com/michellewille2/Virome_Recommendations

Received: 29 May 2024; Accepted: 4 November 2024;
Published online: 20 December 2024

References

1. Zhang, Y. Z., Shi, M. & Holmes, E. C. Using metagenomics to characterize an expanding virosphere. *Cell* **172**, 1168–1172 (2018). Review demonstrating the key role of virome scale studies in expanding the virosphere.
2. Dudas, G. & Batson, J. Accumulated metagenomic studies reveal recent migration, whole genome evolution, and taxonomic incompleteness of orthomyxoviruses. *bioRxiv*, https://doi.org/10.1101/2022.1108.1131.505987 (2022).
3. Parry, R., Wille, M., Turnbull, O. M. H., Geoghegan, J. L. & Holmes, E. C. Divergent influenza-like viruses of amphibians and fish support an ancient evolutionary association. *Viruses* **12**, 1042 (2020).
4. Petrone, M. E. et al. Evidence for an aquatic origin of influenza virus and the order Articulavirales. *bioRxiv*, https://doi.org/10.1101/2023.1102.1115.528772 (2023).
5. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016). **Dramatic expansion of the invertebrate RNA virosphere, redefining our understanding of the host range and evolution of viral families**.
6. Mifsud, J. C. O. et al. Transcriptome mining extends the host range of the Flaviviridae to non-bilaterians. *Virus Evol.* **9**, veac124 (2023).
7. Simmonds, P. et al. ICTV virus taxonomy profile: flaviviridae. *J. Gen. Virol.* **98**, 2–3 (2017).
8. Roux, S. et al. Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019). **Foundational recommendations on presenation of uncultured virus genomes**.
9. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
10. Shaffer, J. P. et al. Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. *Nat. Microbiol.* **7**, 2128–2150 (2022).
11. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).
12. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nat. Rev. Genet.* **20**, 341–355 (2019).
13. Charon, J. et al. Consensus statement from the first RdRp Summit: advancing RNA virus discovery at scale across communities. *Front Virol* 4 (2024). **Outcomes of first RdRp summit, outlining key challenges and solutions for virome-scale research**.
14. Du, Y., Fuhrman, J. A. & Sun, F. ViralCC retrieves complete viral genomes and virus-host pairs from metagenomic Hi-C data. *Nat. Commun.* **14**, 502 (2023).

15. Moshiri, N. ViralConsensus: a fast and memory-efficient tool for calling viral consensus genome sequences directly from read alignment data. *Bioinformatics* **39**, btad317 (2023).
16. Zhou, Z., Martin, C., Kosmopoulos, J. C. & Anantharaman, K. ViWrap: A modular pipeline to identify, bin, classify, and predict viral-host relationships for viruses from metagenomes. *iMeta* https://doi.org/10.1002/imt1002.1118 (2023).
17. Santos, J. D. et al. INSaFLU-TELEVIR: an open web-based bioinformatics suite for viral metagenomic detection and routine genomic surveillance. *Research Square.* https://doi.org/10.21203/rs.21203.rs-3556988/v3556981 (2023).
18. Chen, L. & Banfield, J. F. COBRA improves the completeness and contiguity of viral genomes assembled from metagenomes. *Nat. Microbiol* **9**, 737–750 (2024).
19. Miao, Y. et al. VirGrapher: a graph-based viral identifier for long sequences from metagenomes. *Brief. Bioinform* **25**, bbae036 (2024).
20. Fu, P. et al. VIGA: a one-stop tool for eukaryotic virus identification and genome assembly from next-generation-sequencing data. *Brief. Bioinform* **25**, bbad444 (2023).
21. Tithi, S. S., Aylward, F. O., Jensen, R. V. & Zhang, L. FastViromeExplorer-Novel: Recovering Draft Genomes of Novel Viruses and Phages in Metagenomic Data. *J. Comput Biol.* **30**, 391–408 (2023).
22. Kim, K. et al. VirPipe: an easy-to-use and customizable pipeline for detecting viral genomes from Nanopore sequencing. *Bioinformatics* **39**, btad293 (2023).
23. Wang, X. et al. ViromeFlowX: a comprehensive nextflow-based automated workflow for mining viral genomes from metagenomic sequencing data. *Micro. Genom.* **10**, 001202 (2024).
24. Rangel-Pineros, G. et al. VIRify: An integrated detection, annotation and taxonomic classification pipeline using virus-specific protein profile hidden Markov models. *PLoS Comput Biol.* **19**, e1011422 (2023).
25. Plyusnin, I., Vapalahti, O., Sironen, T., Kant, R. & Smura, T. Enhanced Viral Metagenomics with Lazypipe 2. *Viruses* **15**, 431 (2023).
26. Ru, J., Khan Mirzaei, M., Xue, J., Peng, X. & Deng, L. ViroProfiler: a containerized bioinformatics pipeline for viral metagenomic data analysis. *Gut Microbes* **15**, 2192522 (2023).
27. Song, H., Tithi, S., Aylward, F., Jensen, R. & Zhang, L. Virseqimprover: An Integrated Pipeline for Viral Contig Error Correction, Extension, and Annotation. *Research Square*, https://doi.org/10.21203/rs.21203.rs-3318217/v3318211 (2023).
28. Shen, W. et al. KMCP: accurate metagenomic profiling of both prokaryotic and viral populations by pseudo-mapping. *Bioinformatics* **39**, btac845 (2023).
29. Li, B., Jiao, X. & Liang, G. iVirP: An integrative, efficient, and user-friendly pipeline to annotate viral contigs from raw reads of metagenome or VLP sequencing. *bioRxiv*, https://doi.org/10.1101/2024.1101.1121.576577 (2024).
30. Camargo, A. P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res* **51**, D733–D743 (2023).
31. Kohl, C. et al. Protocol for metagenomic virus detection in clinical specimens. *Emerg. Infect. Dis.* **21**, 48–57 (2015).
32. Chong, R. et al. Fecal viral diversity of captive and wild Tasmanian devils characterised using viron-enriched metagenomics and metatranscriptomics. *J. Virol.* **93**, e00205–e00219 (2019).
33. Yilmaz, P. et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29**, 415–420 (2011). **Critical checklists for descriptions of metadata in virome-scale studies**.
34. Cobbin, J. C., Charon, J., Harvey, E., Holmes, E. C. & Mahar, J. E. Current challenges to virus discovery by meta-transcriptomics. *Curr. Opin. Virol.* **51**, 48–55 (2021). Outlines key challenges to virus discovery.
35. Robbins, A. M. Why scientists should not name diseases based on location. *ASM article*, https://asm.org/Articles/2021/May/Why-Scientists-Should-Not-Name-Diseases-After-Plac (2021).
36. Zerbini, F. M. et al. Differentiating between viruses and virus species by writing their names correctly. *Arch. Virol.* **167**, 1231–1234 (2022).
37. Geoghegan, J. L. & Holmes, E. C. Predicting virus emergence amid evolutionary noise. *Open Biol.* **7**, 170189 (2017).
38. Bergner, L. M. et al. Demographic and environmental drivers of metagenomic viral diversity in vampire bats. *Mol. Ecol.* **29**, 26–39 (2020).
39. Smolak, D. et al. Analysis of RNA virome in rectal swabs of healthy and diarrheic pigs of different age. *Comparative Immunology*. *Microbiol. Infect. Dis.* **90-91**, 101892 (2022).
40. Dominguez-Huerta, G. et al. Diversity and ecological footprint of Global Ocean RNA viruses. *Science* **376**, 1202–1208 (2022).
41. Ettinger, C. L. et al. Highly diverse and unknown viruses may enhance Antarctic endoliths' adaptability. *Microbiome* **11**, 103 (2023).
42. Gregory, A. C. et al. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109 (2019).
43. Lefebvre, M., Theil, S., Ma, Y. X. & Candresse, T. The VirAnnot Pipeline: a resource for automated viral diversity estimation and operational taxonomy units assignation for virome sequencing data. *Phytobiomes J.* **3**, 256–259 (2019).
44. Sachsenroder, J., Twardziok, S. O., Scheuch, M. & Johne, R. The general composition of the faecal virome of pigs depends on age, but not on feeding with a probiotic bacterium. *PLoS ONE* **9**, e88888 (2014).
45. Starr, E. P., Nuccio, E. E., Pett-Ridge, J., Banfield, J. F. & Firestone, M. K. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycline in soil. *PNAS* **116**, 25900–25908 (2019).
46. Zhao, M. et al. Viral metagenomics unveiled extensive communications of viruses within giant pandas and their associated organisms in the same ecosystem. *Sci. Total Environ.* **820**, 153317 (2022).
47. Adriaenssens, E. M. et al. Guidelines for public database submission of uncultivated virus genome sequences for taxonomic classification (vol 41, pg 898, 2023). *Nat. Biotechnol.* **41**, 1346–1346 (2023).
48. Ladner, J. T. et al. Standards for sequencing viral genomes in the era of high-throughput sequencing. *Mbio* **5**, e01360–01314 (2014).
49. Field, D. et al. The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.* **26**, 541–547 (2008).
50. Memish, Z. A. et al. Middle East respiratory syndrome coronavirus in bats, Saudi Arabia. *Emerg. Infect. Dis.* **19**, 1819–1823 (2013).
51. Hill, S. C. et al. Impact of host age on viral and bacterial communities in a waterbird population. *ISME J.* **17**, 215–226 (2023).
52. Abeles, S. R. et al. Human oral viruses are personal, persistent and gender-consistent. *ISME J.* **8**, 1753–1767 (2014).
53. Raghwani, J. et al. Seasonal dynamics of the wild rodent faecal virome. *Mol. Ecol.* https://doi.org/10.1111/mec.16778 (2022).
54. Zhang, W. et al. Virome comparisons in wild-diseased and healthy captive giant pandas. *Microbiome* **5**, 90 (2017).
55. Cao, Z. et al. The gut virome: A new microbiome component in health and disease. *EBioMedicine* **81**, 104113 (2022).
56. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nat. Rev. Genet.* **19**, 514–527 (2021).
57. Mahar, J. E., Shi, M., Hall, R. N., Strive, T. & Holmes, E. C. Comparative analysis of RNA virome composition in rabbits and associated ectoparasites. *J. Virol.* **94**, e02119 (2020).
58. Pettersson, J. H. et al. Circumpolar diversification of the Ixodes uriae tick virome. *PLoS Pathog.* **16**, e1008759 (2020).
59. Asplund, M. et al. Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clin.*

Microbiol. Infect. **25**, 1277–1285 (2019). First description of the reagent viromes, and comprise a critical research to which all virome-scale studies should compare their results.
60. Porter, A. F., Cobbin, J., Li, C. X., Eden, J. S. & Holmes, E. C. Metagenomic identification of viral sequences in laboratory reagents. *Viruses* **13**, 2122 (2021).
61. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
62. Simmonds, P. et al. Four principles to establish a universal virus taxonomy. *Plos Biol.* **21**, e3001922 (2023).
63. Wierenga, J. R. et al. Total infectome investigation of diphtheritic stomatitis in yellow-eyed penguins reveals a novel and abundant megrivirus. *Vet. Microbiol* **286**, 109895 (2023).
64. Costa, V. A. et al. Limited cross-species virus transmission in a spatially restricted coral reef fish community. *Virus Evol.* **9**, vead011 (2023).
65. Qin, J. J. et al. Diversity and potential function of pig gut DNA viruses. *Heliyon* **9**, e14020 (2023).
66. Brito, B. P. et al. Expanding the range of the respiratory infectome in Australian feedlot cattle with and without respiratory disease using metatranscriptomics. *Microbiome* **11**, 158 (2023).
67. Wierenga, J. R. et al. A novel gyrovirus is abundant in yellow-eyed penguin chicks with a fatal respiratory disease. *Virology* **579**, 75–83 (2023).
68. Bustin, S. A. et al. The MIQE Guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.* **55**, 611–622 (2009). MIQE guidelines revolutionised reporting of qPCR experiments to ensure they could be easily interpreted to improve utility by the scientific community.
69. Roux, S. et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *Plos Biol.* **21**, e3002083 (2023).
70. Geoghegan, J. L., Duchene, S. & Holmes, E. C. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog.* **13**, e1006215 (2017).
71. Liu, D., Ma, Y. J., Jiang, X. P. & He, T. T. Predicting virus-host association by Kernelized logistic matrix factorization and similarity network fusion. *Bmc Bioinforma.* **20**, 594 (2019).
72. Babayan, S. A., Orton, R. J. & Streicker, D. G. Predicting reservoir hosts and arthopod vectors from evolutionary signatures in RNA virus genomes. *Science* **362**, 577–580 (2018).
73. Yinda, C. K. et al. Highly diverse population of Picornaviridae and other members of the Picornavirales, in Cameroonian fruit bats. *Bmc Genomics* **18**, 249 (2017).
74. International Committee on Taxonomy of Viruses Executive Committee. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.* **5**, 668–674 (2020).
75. Koonin, E. V. et al. Global organization and proposed megataxonomy of the virus world. *Microbiol Mol. Biol. R.* **84**, e00061–00019 (2020).
76. Dutilh, B. E. et al. Perspective on taxonomic classification of uncultivated viruses. *Curr. Opin. Virol.* **51**, 1–9 (2021).
77. Wille, M., Shi, M., Hurt, A. C., Klaassen, M. & Holmes, E. C. RNA virome abundance and diversity is associated with host age in a bird species. *Virology* **561**, 98–106 (2021).
78. Halabi, K. & Mayrose, I. Mechanisms Underlying Host Range Variation in Flavivirus: From Empirical Knowledge to Predictive Models. *J. Mol. Evolution* **89**, 329–340 (2021).

Acknowledgements

We would like to thank the attendees of the First RdRp Summit, an event focused on establishing an interoperable framework for the discovery, identification and analysis of RNA viruses, who provided feedback and ensured the guidelines proposed are both scientifically sound and practically applicable. We are grateful to E. C. Holmes for his insights.

Author contributions

Conceptualisation – E.H., J.M., M.W. Writing – W.-S.C. E.H., J.M., C.F., M.S., E.S.L., J.L.G., M.W..

Competing interests

The authors declare no competing interests.

Inclusion and ethics

All those who contributed to this manuscript are listed as authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-024-07212-3.

Correspondence and requests for materials should be addressed to Michelle Wille.

Peer review information : *Communications Biology* thanks Pei Hao for their contribution to the peer review of this work. Primary Handling Editor: Tobias Goris.