

Improving the reporting of metagenomic virome-scale data

Wei-Shan Chang^{1,2}, Erin Harvey¹, Jackie Mahar^{1,3}, Jemma L. Geoghegan⁴, Cadhla Firth⁵, Mang Shi⁶, Etienne Simon-Loriere⁷, Edward C. Holmes¹, Michelle Wille^{1,8,*}

¹Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, New South Wales, Australia.

²Health and Biosecurity, Commonwealth Scientific and Industrial Research Organisation, Canberra, Australian Capital Territory, Australia.

³Australian Animal Health Laboratory, Commonwealth Scientific and Industrial Research Organisation, Geelong, Victoria, Australia.

⁴Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand; Institute of Environmental Science and Research, Wellington, New Zealand.

⁵EcoHealth Alliance, New York, USA.

⁶Sun Yat-Sen University, Shenzhen campus of Sun Yat-Sen University, Shenzhen, China

⁷Evolutionary Genomics of RNA Viruses, Institut Pasteur, Université Paris Cité, Paris, France

⁸Centre for Pathogen Genomics, Department of Microbiology and Immunology, at the Peter Doherty Institute for Infection and Immunity. The University of Melbourne, Melbourne, Victoria, Australia.

Corresponding author: michelle.wille@unimelb.edu.au

35 **Abstract**

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

50

51

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 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, the use metagenomic sequencing has substantially expanded our understanding of the host range of virus families. For example, the *Orthomyxoviridae* ^{2, 3, 4, 5} and *Flaviviridae* ^{6, 7}, which were classically defined as mammalian-infecting viral families, are now known to infect hosts across the breadth of the tree of life, including invertebrate phyla.

With decreasing sequencing costs, increasing power of computational resources, and the rapid 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 (Figure 1). Here, we refer to metagenomics as the use of unbiased, high-throughput sequencing of the total genetic material within a sample, in which metatranscriptomics is the use of this technology for the sequencing of RNA specifically. 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 30 in 2011 to 69 in 2021 (based on a keyword search of "virome" on pubmed), (Figure 1) and continues to increase. Associated with this increase are the approximately 750,000 uncultivated viral genomes identified in metagenomic data sets between 2016-2018⁸ and a 6 fold increase in the number of novel virus sequences added to GenBank from 2011 (n= 1,261) to 2021 (n=7,933) (Figure 1). As the cost of sequencing continues to decrease, these numbers will likely continue to rise apace in the coming years.

Compared to studies of the microbiome, or bacterial communities, the integration of metagenomics into virology research remains in its infancy. Microbiome research was arguably 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 also drove innovation in development of tools and technology to facilitate microbiome studies beyond 16S ^{9, 10, 11}. While arguably still the gold 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.

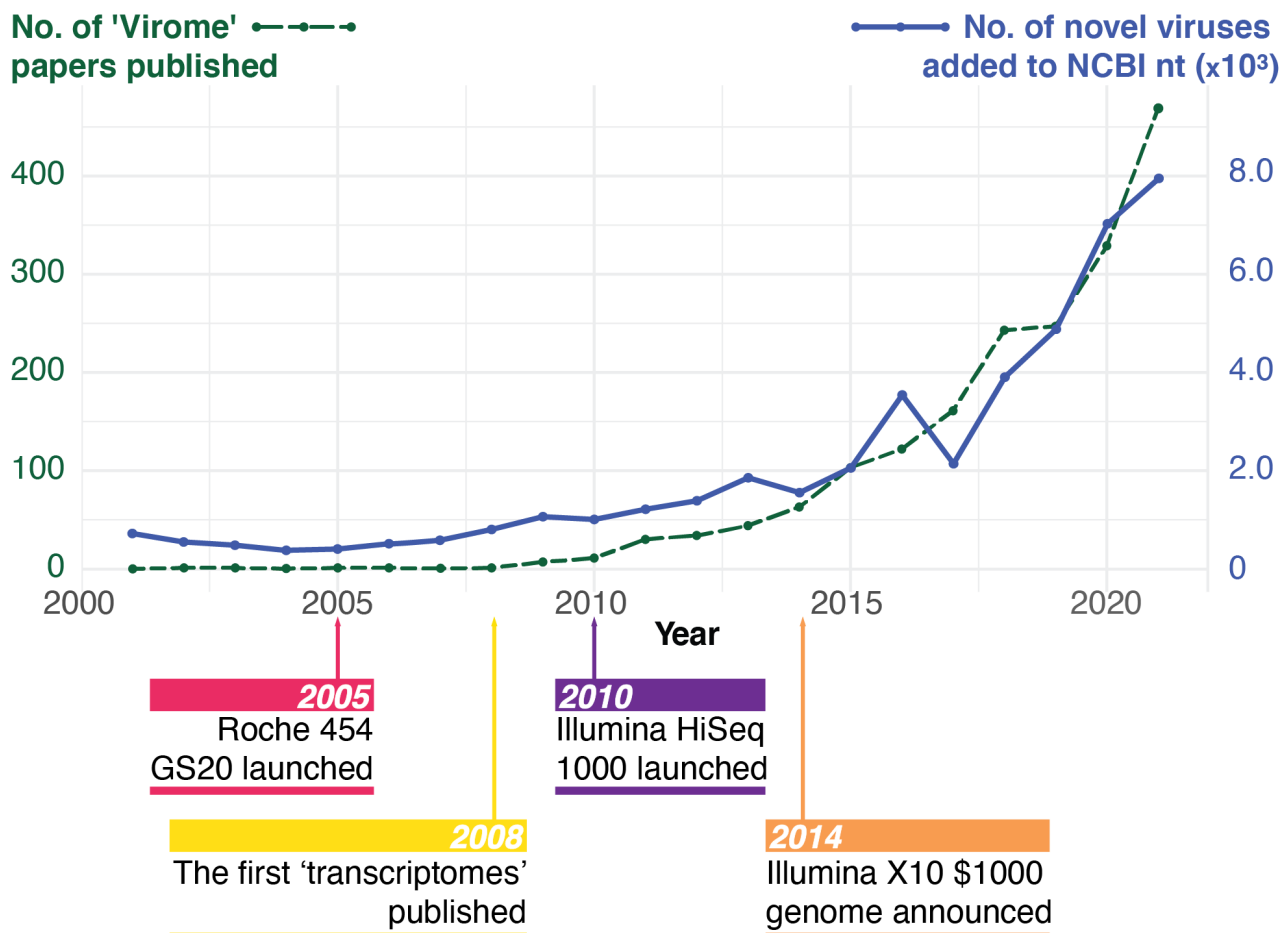


Figure 1: Rapid expansion of metagenomic-based virome studies and novel viral sequences over time. In green: The number of studies published in NCBI's PubMed database each year from 2001 to 2021 that report metagenomic virus discovery/virome analyses [Search query: (metagenomic OR metatranscriptomic) AND (virus OR virome)]. In blue: The number of new virus organisms published in NCBI's nucleotide database each year from 2001 to 2021, sorted by species name. Below the graph, key events in the development of metagenomics are indicated.

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. Indeed, more than 10 new pipelines and packages for virus discovery have

106 been reported in the last two years (e.g.^{13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23}), excluding custom
107 approaches. While this diversification is expected in a growing field and will lead to
108 methodological improvements, the lack of standardized approaches and an absence of
109 appropriately detailed reporting also limits the ability to compare and replicate studies,
110 potentially decreasing their value to the scientific community. To address these
111 deficiencies a systematic approach to data collection, reporting, and analysis is clearly
112 required in the field of viral metagenomics.

113

114 There are a number of shortcomings in many current studies in virus metagenomics.
115 Methods sections commonly lack the detail required for reproducibility, contextualization
116 and evaluation. There are a variety of approaches to sample preparation and data analysis
117 that may be adopted depending on sample type and the specific aims of the project. The
118 methods applied at certain steps will invariably impact the outcome. For example, the use
119 and type of viral enrichment techniques (e.g. particle filtration, nuclease digestion) varies
120 widely, or may not be performed at all. Different extraction kits will similarly alter the
121 detectability and abundance of different types of viruses depending on the methods used
122^{24, 25}. Another common laboratory practice is pooling multiple samples prior to sequencing,
123 yet the description of the pooling strategy is too often described in insufficient detail to be
124 repeated. This also applies to bioinformatic workflows for sequence analysis. For example,
125 some pipelines assemble only sequences (*i.e.* 'reads') that have been identified as viral
126 through sequence similarity searches like BLAST, while others will assemble all reads
127 prior to sequence identification. This choice severely impacts the assembly of highly
128 divergent viruses, biasing downstream estimates of viral diversity, community composition,
129 etc., that should ideally be comparable across studies. The methodological approach to
130 estimating viral abundance from read counts will have similarly important effects on
131 downstream ecological analyses. While it may be theoretically possible to account for
132 some variation in data collection and analytical approaches when performing cross-study
133 comparisons, this is currently impractical due to a lack of transparent and detailed
134 methodological reporting.

135

136 Insufficient detail in the sharing of metagenomic data, associated metadata, and the
137 reporting of analytical results is also commonplace in virome-scale studies. For example,
138 despite existing checklists²⁶, accompanying metadata (e.g., collection date, location, host,
139 sample type, disease state) may be excluded or not comprehensive, reducing the value of
140 these data and the ability to place it in the correct ecological or evolutionary context. The

141 use of raw data outputs from automated bioinformatics pipelines has led to the publication
142 of virus discovery studies that provide limited information on the virus beyond its
143 unannotated genomic sequence. This transfers the burden of quality control, annotation,
144 and taxonomic assignment upon database curators and/or researchers using these data in
145 future virus discovery efforts. This can be resolved through sequence annotation and the
146 inclusion of phylogenies as discussed below.

147

148 A key challenge in viral metagenomics is the correct association of a viral sequence with
149 its host; however, host-virus associations are sometimes neglected altogether or
150 incorrectly reported. This may be because the host of a particular virus is not necessarily
151 the species from which it was sampled, as many viruses in a metagenomic sample
152 originate from the sampled species' microbiome, diet, or result from laboratory
153 contamination ²⁷. This is described in more detail in Box 1. Determining host associations
154 is made more complex through improper naming of published viral sequences. Despite
155 being inconsistent with best practice or ICTV guidelines ^{28, 29} (Box 2), the name of the
156 sampled organism is often included in novel virus names which can be misleading when
157 the sampled species has not been determined as the definitive host. For example, neither
158 Bat Iflavirus (GenBank Accession NC_033823) nor Goose Dicistrovirus (GenBank
159 Accession NC_029052) have reservoirs in vertebrates. Rather, these viruses are likely
160 associated with invertebrates and hence comprise the viruses of the diet of the sampled
161 vertebrate hosts. Erroneous host associations in public databases can lead to cascades of
162 host mischaracterization, and have the potential to result in incorrect evolutionary or
163 ecological inferences.

164

165 Variation in the approaches used for metagenomic data analysis is equally problematic for
166 the interpretation of virome data, particularly when limited methodological detail is
167 provided. As viruses do not possess a universally conserved gene for taxonomic
168 assignment akin to the 16S rRNA gene used in microbiome studies ³⁰, and because such
169 a large proportion of the virosphere remains unresolved ³¹, virome characterisation is often
170 complex and requires thorough and supervised analyses. Inadequate analyses can lead to
171 both spurious and redundant results. For example, many virome studies report viruses
172 without conducting phylogenetic analyses, a vital step for virus classification and the
173 baseline for many evolutionary and ecological inferences (*e.g.* ^{32, 33}). Providing viral gene
174 sequences can be reliably aligned, phylogenetics is arguably the best way to validate
175 novel viral sequences and determine their taxonomy, while also providing clues to the

likely host or whether the virus may be a contaminant²⁷ (see Box 3). Yet this step is frequently 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 through sequence similarity alone, which are problematic when performed without contig verification such as virus identification or sequence annotation (e.g.^{32, 34, 35, 36, 37, 38}). 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^{39, 40}. As the 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, deficiencies in reporting results and methodologies may limit the value of time-intensive and costly metagenomic studies to the scientific community. In worst cases, the deposition of inappropriately characterised sequence data into public databases, may 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 shortcomings

Without specific guidelines, most genome sequences in databases are sparsely annotated with the information required to guide data interpretation and knowledge generation²⁶. 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.genesc.org/pages/standards/checklists.html>). An abbreviated summary of the checklist relevant to the data produced in virome-scale studies is presented in Figure 2 and includes the Minimum information about a marker gene sequence (MIMARKS) checklist as an extension to the Minimum Information about any Sequence (MIxS) list²⁶, the Minimum Information to report Uncultured Virus Genome

(MIUViG)⁸, and recommendations presented in Ladner *et al.* (2014)⁴¹. These checklists provide an ideal starting point for developing a comprehensive set of recommendations for the presentation of virome-scale data analysis and the resulting genomes.

Briefly, MlxS 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^{26, 42}. A useful extension of the MlxS checklist is the MIMARKS checklist²⁶. 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 (Figure 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.* 2017) and MIUViG⁸. The former is targeted specifically toward bacterial genomes, while the MIUViG checklist, particularly when combined with the recommendations of Ladner *et al.* (2014)⁴¹, 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, 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 (Figure 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 (Figure 3).

	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

Figure 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.* (2011)²⁶, those from MIUVIG are those outlined in Roux *et al.* (2019)⁸, and those included in a standard virus genome are those outlined in Ladner *et al.* (2014)⁴¹. (Abbreviations -- 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)

10 key recommendations for reporting virome-scale studies

- Sample collection, storage, transport, and metadata.** Complete information is required on materials used for the collection and storage of samples (e.g., type of swab, transport media, tube, etc.), as each can have important consequences for nucleic acid quality⁴³. Sample metadata should 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. For example, age⁴⁴, sex⁴⁵, season⁴⁶, disease status⁴⁷, and phenotypic characteristics^{48, 49} all have the potential to influence the virome. Detailed metadata checklists presented in Yilmaz *et al.* (2011)²⁶ should be consistently considered.

- 268 2. **Sample preparation and viral enrichment or depletion protocols.** Details and
269 timing of nucleic acid extraction methods, virus enrichment, amplification, or
270 depletion protocols, sequence library preparation, negative/positive controls should
271 be presented. A detailed description of the approach used for sample pooling
272 should be presented if relevant, including the number of samples/pool and method
273 of pooling (by volume, concentration, molarity, etc.) Sample preparation
274 approaches, such as pooling, can vastly affect the interpretation of results, such as
275 when calculating viral sequence abundance and richness.
276
- 277 3. **Sequencing methodology.** A description of the sequencing methodology,
278 including platform, read length, and whether paired- or single-end is required.
279 Critically, this should also include results on the number of sequences generated
280 per library.
281
- 282 4. **Bioinformatic approaches.** Bioinformatic pipelines should be fully described and
283 reproduceable. At a minimum, all details around software, version, parameter
284 settings and manual steps should be described. Ideally, all newly generated or
285 custom code should be deposited in open-source repositories such as GitHub.
286 Details regarding quality control, trimming, assembly, contig annotation, and read
287 mapping should be reported. The bioinformatic approaches used for taxonomic
288 assignment of contigs (or reads) should be specified, including the database
289 queried. The number of contigs and contig lengths (range) should be reported. For
290 contigs of interest (*i.e.* those comprising viruses), the results of sequence similarity
291 searches (*e.g.* BLAST) including closest genetic relative, percent sequence
292 identities, alignment lengths, e-values, and contig lengths should also be provided
293 as supplementary data (if not in the main text).
294
- 295 5. **Methodological checks and balances.** “Index hopping” (wherein a proportion of
296 reads are incorrectly indexed, usually 0.01-0.1% of reads if using common Illumina
297 technologies) should be accounted for during data analysis and details on how
298 index-hopping was addressed should be reported (*e.g.*^{50, 51}). Efforts should be
299 made to confirm that viral contigs were not derived from reagents or incidental
300 contaminants. This can be achieved by comparing the results to lists of known
301 reagent contaminants^{52, 53} as well as to experiment-specific no-template (negative)
302 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.** At a minimum, 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 genome organisation is different than expected, authors should provide details on how they confirmed the sequence. 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. ⁵⁴).

7. **Phylogenetic analysis of putative viral transcripts.** Phylogenetic analysis of newly identified viral transcripts should be considered 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 crucial to include other members of the genus, family, or order to provide adequate context (expanded upon in Box 3). Phylogenetic trees should be estimated with care: alignments should be made using algorithms that can adequately deal with large gaps in alignments (and manually checked), and trees estimated using robust methods. The genomic region or protein used, alignment length, and tools used for sequence alignment should be reported, along with methods used for the removal of poorly aligned regions, model testing, tree inference, and nodal support estimates (e.g. bootstrapping). Ideally, data underlying trees should be made

available through open-source, persistently identified repositories such as Zenodo or FigShare.

8. **Presenting putatively novel viruses.** When considering assigning taxonomy to newly characterized viruses, thresholds of nucleotide and/or amino acid similarity should be provided. Preferentially, virus family-specific guidelines developed by the ICTV should be followed. The ICTV criteria for the demarcation of viral species is 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 virus names proposed by the study authors would constitute the sequence or virus name; only the ICTV designates species and associated species names (Box 2). 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 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 whereby, for example, a disease is incorrectly attributed to a novel virus detection and not biologically relevant. At a minimum, phylogenetic analysis (as discussed in 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.* (2021)²⁷, and in Box 1.
10. **Data sharing principles: Findable, Accessible, Interoperable, Reusable.** At a minimum, sequencing reads, including negative control libraries, 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) and must be

linked to the deposited sequencing libraries. ORF translations should also be included in the GenBank/ENA record for the sequence to be included in NCBI protein database. Custom code or newly developed bioinformatic approaches or pipelines used for data analysis should be freely available on open-source platforms such as GitHub. Ideally, laboratory protocols or workflows should be uploaded to repositories that provide persistent identifiers (e.g. DOIs) such as protocols.io, Critically, persistent, 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.

Sample collection, preparation, sequencing	Contig assembly and identification	Virus annotation and confirmation	Virus presentation
① METADATA <ul style="list-style-type: none"> - materials for collection and storage: <ul style="list-style-type: none"> > swab, media, tubes - sample metadata: <ul style="list-style-type: none"> > sample type, location, date, host, age, sex, disease, etc - checklists from Yilmaz <i>et al.</i> ② SAMPLE PREP <ul style="list-style-type: none"> - pooling approach if used - extraction methods - enrichment, amplification or depletion methods - negative/positive controls - library preparation ③ SEQUENCING <ul style="list-style-type: none"> - sequencing platform - paired or single reads - number of reads per library 	④ BIOINFORMATIC APPROACHES <ul style="list-style-type: none"> - pipelines fully described: <ul style="list-style-type: none"> > QC, trimming, assembly, annotation, read mapping tools and parameters - number/length of contigs assembled - for viral contigs: <ul style="list-style-type: none"> > Top hit name & accession > % identity > evaluate ⑤ CHECKS & BALANCES <ul style="list-style-type: none"> - index hopping addressed - reagent contaminant checked - misassembly checks 	⑥ ANNOTATION <ul style="list-style-type: none"> - ORFs found and verified - other annotations considered, including: <ul style="list-style-type: none"> > domains, motifs, mature peptides, IRES sites, etc - verify EVEs ⑦ PHYLOGENETICS <ul style="list-style-type: none"> - alignment method and parameters - alignment length, region used - how gaps treated - model testing performed - phylogenetic tree software and parameters (e.g. bootstrap approach) - more details in Box 3 	⑧ VIRUS DESCRIPTION <ul style="list-style-type: none"> - assessment criteria provided - unique virus name if novel, more details on naming in Box 2 - summary including: <ul style="list-style-type: none"> > genome length, completeness, number of segments, etc > Link to GenBank record and metadata ⑨ HOST ASSOCIATION <ul style="list-style-type: none"> - identify viruses of biological relevance vs viruses of diet, microbiome - more details in Box 1 ⑩ DATA SHARING <ul style="list-style-type: none"> - Findable, Accessible, Interoperable, Reusable

Figure 3. Summary of data presentation features we propose for inclusion in all virome studies. A tabular checklist is provided as supplemental Table S1.

Community use of proposed guidelines

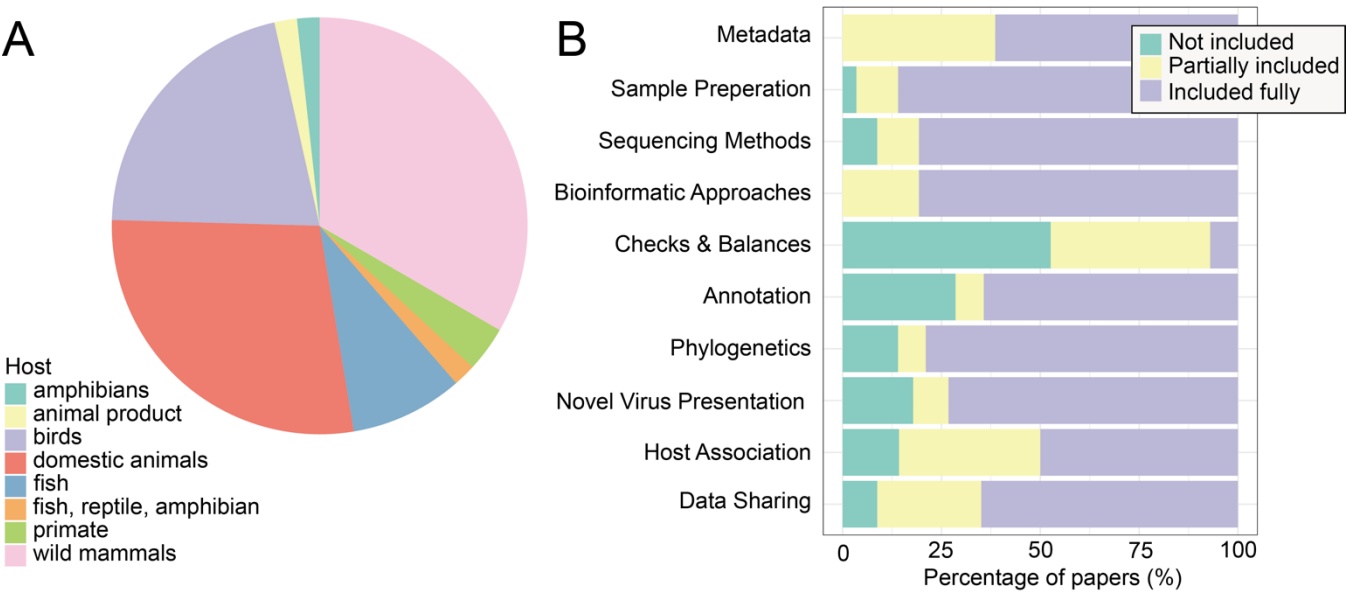
We have provided a road map for metagenomic virome-scale data reporting through robust recommendations that build on components already reported in a substantial proportion of studies, with key foundations in available minimum standards checklists. Importantly, the data reporting road map provided here 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 2021, focussing on vertebrate animal systems. Specifically, we identified studies focussing on non-human vertebrates (n=36) in all PubMed hits for “virome” (n=471). We supplemented this with an additional 21 studies that were manually identified, for a total of 57 studies. Overall, we found that most studies included details of sample preparation,

399 sequencing methods and bioinformatic approaches, either in detail or partially (Figure 4).
400 Across the 10 recommendations, we found the lowest uptake was on “checks and
401 balances” (recommendation 5), which comprises the inclusion of no template control
402 libraries to identify putative reagent contamination, and addresses index hopping.
403 However, as the new field of viral metagenomics matures, so too will our appreciation of
404 the limitations of the associated tools and techniques, and as a result, more checks and
405 balances will be incorporated. As such, the current low uptake of this recommendation is
406 most likely a reflection of an area where there is the largest capacity for improvement. Viral
407 annotation was also found to have substantial room for improvement, with more than a
408 quarter of studies failing to include this vital information (Figure 4). Notably, only 4 studies
409 included all items fully, demonstrating the need and opportunity for improvement moving
410 forward ^{55, 56, 57, 58}.

411
412 The current inconsistency in methods and results reporting is most likely the direct result of
413 a lack of recommendations available in this rapidly expanding field. We anticipate that the
414 unified and inclusive framework we have presented here will be substantially more
415 straightforward and accessible (*i.e.*, if you build it, they will come). As we have harmonised
416 and built upon many pieces of existing information with accepted value in the community,
417 we expect high uptake across the field, which will lead to a substantial improvement in the
418 utility of virome-scale metagenomic research.

419



420

421

Figure 4. Papers published in 2021 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.

Conclusion

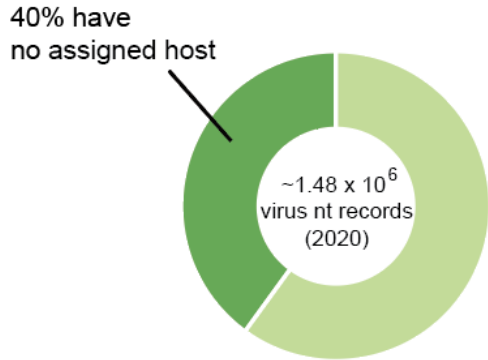
There is a lack of consensus on how best to perform virome-scale metagenomic research. The problem is exacerbated by a lack of sufficient experimental detail in many publications, which impedes a reader's ability to critically evaluate the quality of the results presented, to repeat the experiments, or to utilise the published results in their own research. We have provided a set of 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 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. As such, the development of minimum standards and guidelines is critical. 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 qPCR⁵⁹. 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.

BOX 1. Host association

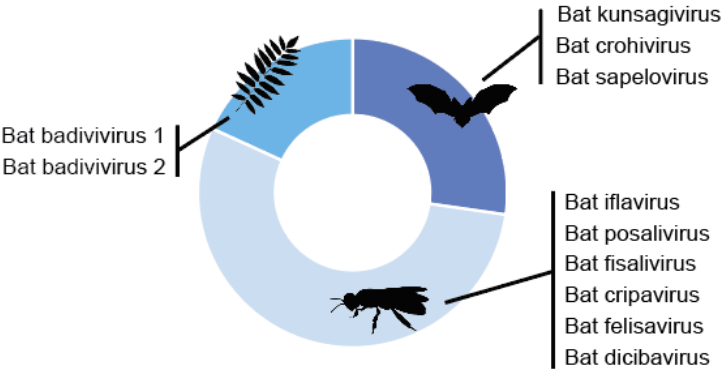
Clarifying potential host associations is of critical importance to resolving the virosphere, and at a lower level, revealing viral dynamics and relationships in the context of "host-pathogen" networks. Inaccuracies in host association and/or naming viruses after hosts despite incongruous host-association (Figure 5) leads 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.

458 Identifying host associations is challenging, and a number of different approaches have
 459 been put forward, as summarised in Cobbin *et al.* (2021)²⁷. The most straightforward is to
 460 conduct phylogenetic analysis to identify the host association of closely related viruses,
 461 assuming virus-host codivergence (being mindful that hosts could be misassigned in the
 462 database)⁶⁰. Beyond phylogenetics, a variety of other approaches may be employed
 463 including: (1) exploring signatures of virus and host genome coevolution by comparing the
 464 virus and potential host codon pairing and/or oligonucleotide frequency (ONF) patterns⁶¹.
 465 (2) Correlating viral abundance with the abundance of intra-host microbe marker genes in
 466 cases where metatranscriptomic sequencing has been employed. (3) Conducting large-
 467 scale virus-host association studies in which diverse host data sets from resources such
 468 as the Sequence Read Archive (SRA) are mined for viruses, and (4) excluding or
 469 identifying the presence of other potential host species by analysing non-viral sequences
 470 from the same library. Tools using a combination of approaches are showing utility and
 471 high levels of accuracy. For example, a machine learning model using a combination of
 472 phylogenetics and biases in viral genome composition was successfully used to identify
 473 arthropod vectors for a substantial array of viruses⁶².
 474

A. Proportion virus nt records
 with host association



B. Virus names in public databases
 may have misleading names



475
 476
 477 **Figure 5.** Current state of host association. (A) As indicated in dark orange, a substantial
 478 proportion of viral records in GenBank do not have an associated host. Modified from Cobbin *et al.*
 479 (2021)²⁷. (B) Eleven picornaviruses recovered from bat faeces, all including “bat” in the virus name,
 480 have 3 different hosts as indicated by silhouettes overlaying the plot. Modified from Yinda *et al.*
 481 (2017)⁶³.
 482

483 **BOX 2. Improving clarity in virus presentation**

484 Virus taxonomy and naming is under the purview of the International Committee on the
485 Taxonomy of Viruses (ICTV; <https://ictv.global/>)⁶⁴. Currently we are seeing a substantial
486 and continued overhaul of virus taxonomy and nomenclature, with changes occurring on
487 different timelines among the different subcommittees/virus families (e.g. ⁶⁵). This is
488 creating substantial challenges in presenting both novel and established virus species in
489 scientific articles. Herein we provide some suggestions to improve the clarity of virus
490 names presented in studies, but with the caveat that it's a continually evolving
491 landscape/situation.

492 For clarity only, it is preferable to provide highly divergent virus sequences that potentially
493 constitute new species with a unique virus name. It is important to note that virus names
494 provided by the author are not synonymous with virus species names, which are decided
495 by the ICTV following assessment and ratification of novel viruses⁶⁴. Using contig names
496 or complex coded names are not ideal as they may be impossible to decipher by others
497 wishing to include sequences for comparison in future studies. For example, "par083ade1"
498 or "ERX3854945_k113_22474" are not particularly useful virus names. Current ICTV
499 guidelines indicate that viruses cannot be named after people, locations, host species, or
500 copyright protected names ²⁸. Location and host species should not be included in virus
501 names as the point of detection may not be a true reflection of spatial or host range. It is
502 important to confirm that proposed names have not been used prior.

503 Presentation of previously described viruses should be done in accordance with both the
504 ICTV and field-specific nomenclature. Virus species names and taxonomy should be
505 presented as outlined by Zerbini *et al.* (2022)²⁹. Providing additional detail regarding
506 clades, variants, strains, subtypes or serotypes within established nomenclature systems
507 is crucial for improving the value of the presented findings.

508

509 **BOX 3. Phylogenetic analysis as a key step in virus verification**

510

511 Given most of the virosphere remains undiscovered ³¹, validation and characterisation of
512 novel viral contigs is imperative, and relying on only viral operational taxonomic units
513 (vOTU's), diversity statistics, or BLAST results is not sufficient. To validate novel viral
514 sequences and their relationship to other viruses, a robust phylogenetic analysis is
515 required. Pipelines relying only on BLAST have a severe shortcoming, particularly in the
516 context of highly divergent virus sequences. These divergent virus sequences may have
517 <40% amino acid similarity, and therefore the closest relative identified by BLAST is highly
518 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 RdRp. Additionally, when utilizing methods such as maximum likelihood with the incorporation of an appropriate nucleotide or amino acid substitution model, evolutionary relationships can be assessed. 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*⁵⁸. 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 interferences 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, 66}. 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^{52, 53}. 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.

Competing Interests

We declare that none of the authors have competing financial or non-financial interests.

554

555 Inclusion and Ethics

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

557

558 Data availability

559 No new data were generated in this study

560

561 Code availability

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

563

564 References

565

- 566 1. Zhang, Y.Z., Shi, M. & Holmes, E.C. Using metagenomics to characterize an
567 expanding virosphere. *Cell* **172**, 1168-1172 (2018).
- 568 2. Dudas, G. & Batson, J. Accumulated metagenomic studies reveal recent migration,
569 whole genome evolution, and taxonomic incompleteness of orthomyxoviruses.
570 *bioRxiv*, doi: <https://doi.org/10.1101/2022.1108.1131.505987> (2022).
- 571 3. Parry, R., Wille, M., Turnbull, O.M.H., Geoghegan, J.L. & Holmes, E.C. Divergent
572 influenza-like viruses of amphibians and fish support an ancient evolutionary
573 association. *Viruses* **12**, doi: 10.3390/v12091042 (2020).
- 574 4. Petrone, M.E. *et al.* Evidence for an aquatic origin of influenza virus and the order
575 Articulavirales. *bioRxiv*, doi: <https://doi.org/10.1101/2023.1102.1115.528772> (2023).
- 576 5. Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature* **540**, 539-543
577 (2016).
- 578 6. Mifsud, J.C.O. *et al.* Transcriptome mining extends the host range of the
579 Flaviviridae to non-bilaterians. *Virus Evol* **9**, veac124. doi: 110.1093/ve/veac1124
580 (2023).
- 581 7. Simmonds, P. *et al.* ICTV Virus Taxonomy Profile: Flaviviridae. *Journal of General*
582 *Virology* **98**, 2-3 (2017).
- 583 8. Roux, S. *et al.* Minimum information about an uncultivated virus genome (MIUViG).
584 *Nat Biotechnol* **37**, 29-37 (2019).
- 585 9. Thompson, L.R. *et al.* A communal catalogue reveals Earth's multiscale microbial
586 diversity. *Nature* **551**, 457-463 (2017).
- 587 10. Shaffer, J.P. *et al.* Standardized multi-omics of Earth's microbiomes reveals
588 microbial and metabolite diversity. *Nat Microbiol* **7**, 2128-2150 (2022).
- 589 11. Olm, M.R. *et al.* inStrain profiles population microdiversity from metagenomic data
590 and sensitively detects shared microbial strains. *Nat Biotechnol* **39**, 727-736 (2021).
- 591 12. Chiu, C.Y. & Miller, S.A. Clinical metagenomics. *Nature Reviews Genetics* **20**, 341-
592 355 (2019).
- 593 13. Budkina, A.Y. *et al.* Utilizing the VirIdAI pipeline to search for viruses in the
594 metagenomic data of bat samples. *Viruses* **13**, 2006, doi: 10.3390/v13102006
595 (2021).
- 596 14. Camargo, A.P. *et al.* IMG/VR v4: an expanded database of uncultivated virus
597 genomes within a framework of extensive functional, taxonomic, and ecological
598 metadata. *Nucleic Acids Res*, doi: 10.1093/nar/gkac1037 (2022).

- 599 15. de Vries, J.J.C. *et al.* Benchmark of thirteen bioinformatic pipelines for
600 metagenomic virus diagnostics using datasets from clinical samples. *Journal of*
601 *Clinical Virology* **141**, 104908. doi: 104910.101016/j.jcv.102021.104908 (2021).
- 602 16. Du, Y., Fuhrman, J.A. & Sun, F. ViralCC retrieves complete viral genomes and
603 virus-host pairs from metagenomic Hi-C data. *Nature Communications* **14**, 502. doi:
604 510.1038/s41467-41023-35945-y (2023).
- 605 17. Mastriani, E., Bienes, K.M., Wong, G.R.Y. & Berthet, N. PIMGAVir and Vir-MinION:
606 two viral metagenomic pipelines for complete baseline analysis of 2nd and 3rd
607 generation data. *Viruses* **14**, 1260. doi: 1210.3390/v14061260 (2022).
- 608 18. Moshiri, N. ViralConsensus: a fast and memory-efficient tool for calling viral
609 consensus genome sequences directly from read alignment data. *Bioinformatics* **39**,
610 btad317. doi: 310.1093/bioinformatics/btad1317 (2023).
- 611 19. Perot, P., Bigot, T., Temmam, S., Regnault, B. & Eloit, M. Microseek: a protein-
612 based metagenomic pipeline for virus diagnostic and discovery. *Viruses* **14**, doi:
613 10.3390/v14091990 (2022).
- 614 20. Plyusnin, I. *et al.* Novel NGS pipeline for virus discovery from a wide spectrum of
615 hosts and sample types. *Virus Evol* **6**, veaa091. doi: 010.1093/ve/veaa1091 (2020).
- 616 21. Rangel-Pineros, G. *et al.* VIRify: an integrated detection, annotation and taxonomic
617 classification pipeline using virus-specific protein profile hidden Markov models.
618 *PLoS Computational Biology* **19**, e1011422.
619 <https://doi.org/10.1011371/journal.pcbi.1011422> (2022).
- 620 22. Wylie, T.N. & Wylie, K.M. ViroMatch: a computational pipeline for the detection of
621 viral sequences from complex metagenomic data. *Microbiol Resour Ann* **10**,
622 e01468-01420. doi: 01410.01128/MRA.01468-01420 (2021).
- 623 23. Zhou, Z., Martin, C., Kosmopoulos, J.C. & Anantharaman, K. ViWrap: A modular
624 pipeline to identify, bin, classify, and predict viral-host relationships for viruses from
625 metagenomes. *iMeta*, doi: <https://doi.org/10.1002/imt1002.1118> (2023).
- 626 24. Kohl, C. *et al.* Protocol for metagenomic virus detection in clinical specimens.
627 *Emerg Infect Dis* **21**, 48-57 (2015).
- 628 25. Chong, R. *et al.* Fecal viral diversity of captive and wild Tasmanian devils
629 characterised using viron-enriched metagenomics and metatranscriptomics. *J Virol*
630 **93**, e00205-00219. doi: 00210.01128/JVI.00205-00219 (2019).
- 631 26. Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS)
632 and minimum information about any (x) sequence (MIXS) specifications. *Nat*
633 *Biotechnol* **29**, 415-420 (2011).
- 634 27. Cobbin, J.C., Charon, J., Harvey, E., Holmes, E.C. & Mahar, J.E. Current
635 challenges to virus discovery by meta-transcriptomics. *Current Opinion in Virology*
636 **51**, 48-55 (2021).
- 637 28. Robbins, A.M. Why scientists should not name diseases based on location. *ASM*
638 *article*, [https://asm.org/Articles/2021/May/Why-Scientists-Should-Not-Name-](https://asm.org/Articles/2021/May/Why-Scientists-Should-Not-Name-Diseases-After-Place)
639 [Diseases-After-Place](https://asm.org/Articles/2021/May/Why-Scientists-Should-Not-Name-Diseases-After-Place) (2021)
- 640 29. Zerbini, F.M. *et al.* Differentiating between viruses and virus species by writing their
641 names correctly. *Arch Virol* **167**, 1231-1234 (2022).
- 642 30. Wensel, C.R., Pluznick, J.L., Salzberg, S.L. & Sears, C.L. Next-generation
643 sequencing: insights to advance clinical investigations of the microbiome. *J Clin*
644 *Invest* **132**, e154944. doi: 154910.151172/JCI154944 (2022).
- 645 31. Geoghegan, J.L. & Holmes, E.C. Predicting virus emergence amid evolutionary
646 noise. *Open Biol* **7**, 170189. doi:170110.171098/rsob.170189 (2017).
- 647 32. Bergner, L.M. *et al.* Demographic and environmental drivers of metagenomic viral
648 diversity in vampire bats. *Molecular Ecology* **29**, 26-39 (2020).

- 649 33. Smolak, D. *et al.* Analysis of RNA virome in rectal swabs of healthy and diarrheic
650 pigs of different age. *Comparative Immunology, Microbiology & Infectious Diseases*
651 **90-91**, 101892 (2022)
- 652 34. Dominguez-Huerta, G. *et al.* Diversity and ecological footprint of Global Ocean RNA
653 viruses. *Science* **376**, 1202-1208 (2022).
- 654 35. Ettinger, C.L. *et al.* Highly diverse and unknown viruses may enhance Antarctic
655 endoliths' adaptability. *Microbiome* **11**, 103. doi: 110.1186/s40168-40023-01554-
656 40166 (2023).
- 657 36. Gregory, A.C. *et al.* Marine DNA viral macro- and microdiversity from pole to pole.
658 *Cell* **177**, 1109. doi: 1110.1016/j.cell.2019.1103.1040 (2019).
- 659 37. Lefebvre, M., Theil, S., Ma, Y.X. & Candresse, T. The VirAnnot Pipeline: a resource
660 for automated viral diversity estimation and operational taxonomy units assignment
661 for virome sequencing data. *Phytobiomes J* **3**, 256-259 (2019).
- 662 38. Sachsenroder, J., Twardziok, S.O., Scheuch, M. & Johne, R. The general
663 composition of the faecal virome of pigs depends on age, but not on feeding with a
664 probiotic bacterium. *PLoS One* **9**, e88888. doi: 88810.81371/journal.pone.0088888
665 (2014).
- 666 39. Starr, E.P., Nuccio, E.E., Pett-Ridge, J., Banfield, J.F. & Firestone, M.K.
667 Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape
668 carbon cycle in soil. *PNAS* **116**, 25900-25908 (2019).
- 669 40. Zhao, M. *et al.* Viral metagenomics unveiled extensive communications of viruses
670 within giant pandas and their associated organisms in the same ecosystem. *Sci*
671 *Total Environ* **820**, 153317. doi: 153310.151016/j.scitotenv.152022.153317 (2022).
- 672 41. Ladner, J.T. *et al.* Standards for sequencing viral genomes in the era of high-
673 throughput sequencing. *Mbio* **5**, e01360-01314. doi: 01310.01128/mBio.01360-
674 01314 (2014).
- 675 42. Field, D. *et al.* The minimum information about a genome sequence (MIGS)
676 specification. *Nat Biotechnol* **26**, 541-547 (2008).
- 677 43. Memish, Z.A. *et al.* Middle East respiratory syndrome coronavirus in bats, Saudi
678 Arabia. *Emerg Infect Dis* **19**, 1819-1823 (2013).
- 679 44. Hill, S.C. *et al.* Impact of host age on viral and bacterial communities in a waterbird
680 population. *ISME Journal* **17**, 215-226 (2023).
- 681 45. Abeles, S.R. *et al.* Human oral viruses are personal, persistent and gender-
682 consistent. *ISME Journal* **8**, 1753-1767 (2014).
- 683 46. Raghwan, J. *et al.* Seasonal dynamics of the wild rodent faecal virome. *Molecular*
684 *Ecology*, doi: 10.1111/mec.16778 (2022).
- 685 47. Zhang, W. *et al.* Virome comparisons in wild-diseased and healthy captive giant
686 pandas. *Microbiome* **5**, 90. doi: 10.1186/s40168-40017-40308-40160 (2017).
- 687 48. Cao, Z. *et al.* The gut virome: A new microbiome component in health and disease.
688 *EBioMedicine* **81**, 104113. doi: 104110.101016/j.ebiom.102022.104113 (2022).
- 689 49. Liang, G. & Bushman, F.D. The human virome: assembly, composition and host
690 interactions. *Nature Reviews Genetics* **19**, 514-527 (2021).
- 691 50. Mahar, J.E., Shi, M., Hall, R.N., Strive, T. & Holmes, E.C. Comparative analysis of
692 RNA virome composition in rabbits and associated ectoparasites. *J Virol* **94**,
693 e02119-02119. doi: 02110.01128/JVI.02119-02119 (2020).
- 694 51. Pettersson, J.H. *et al.* Circumpolar diversification of the Ixodes uriae tick virome.
695 *PLoS Pathogens* **16**, e1008759. doi: 1008710.1001371/journal.ppat.1008759
696 (2020).
- 697 52. Asplund, M. *et al.* Contaminating viral sequences in high-throughput sequencing
698 viromics: a linkage study of 700 sequencing libraries. *Clinical Microbiology and*
699 *Infection* **25**, 1277-1285 (2019).

- 700 53. Porter, A.F., Cobbin, J., Li, C.X., Eden, J.S. & Holmes, E.C. Metagenomic
701 identification of viral sequences in laboratory reagents. *Viruses* **13**, doi:
702 10.3390/v13112122 (2021).
- 703 54. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-
704 assembled viral genomes. *Nat Biotechnol* **39**, 578-585 (2021).
- 705 55. Geoghegan, J.L. *et al.* Virome composition in marine fish revealed by meta-
706 transcriptomics. *Virus Evol* **7**, veab005. doi: 010.1093/ve/veab1005 (2021).
- 707 56. Ning, S. *et al.* Virome in fecal samples from wild Giant Pandas (*Ailuropoda*
708 *melanoleuca*). *Frontiers in Veterinary Science* **8**, 767494. doi:
709 767410.763389/fvets.762021.767494 (2021).
- 710 57. Costa, V.A. *et al.* Metagenomic sequencing reveals a lack of virus exchange
711 between native and invasive freshwater fish across the Murray-Darling Basin,
712 Australia. *Virus Evol* **7**, doi: 10.1093/ve/veab1034 (2021).
- 713 58. Wille, M., Shi, M., Hurt, A.C., Klaassen, M. & Holmes, E.C. RNA virome abundance
714 and diversity is associated with host age in a bird species. *Virology* **561**, 98-106
715 (2021).
- 716 59. Bustin, S.A. *et al.* The MIQE Guidelines: minimum information for publication of
717 quantitative real-time PCR experiments. *Clin Chem* **55**, 611-622 (2009).
- 718 60. Geoghegan, J.L., Duchene, S. & Holmes, E.C. Comparative analysis estimates the
719 relative frequencies of co-divergence and cross-species transmission within viral
720 families. *PLoS Pathogens* **13**, e1006215. doi:
721 1006210.1001371/journal.ppat.1006215 (2017)
- 722 61. Liu, D., Ma, Y.J., Jiang, X.P. & He, T.T. Predicting virus-host association by
723 Kernelized logistic matrix factorization and similarity network fusion. *Bmc*
724 *Bioinformatics* **20**, 594. doi: 510.1186/s12859-12019-13082-12850 (2019).
- 725 62. Babayan, S.A., Orton, R.J. & Streicker, D.G. Predicting reservoir hosts and
726 arthropod vectors from evolutionary signatures in RNA virus genomes. *Science* **362**,
727 577-580 (2018).
- 728 63. Yinda, C.K. *et al.* Highly diverse population of Picornaviridae and other members of
729 the Picornavirales, in Cameroonian fruit bats. *Bmc Genomics* **18**, 249. doi:
730 210.1186/s12864-12017-13632-12867 (2017).
- 731 64. International Committee on Taxonomy of Viruses Executive Committee. The new
732 scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks.
733 *Nature Microbiology* **5**, 668-674 (2020).
- 734 65. Koonin, E.V. *et al.* Global organization and proposed megataxonomy of the virus
735 world. *Microbiol Mol Biol R* **84**, e00061-00019. doi: 00010.01128/MMBR.00061-
736 00019 (2020).
- 737 66. Halabi, K. & Mayrose, I. Mechanisms Underlying Host Range Variation in Flavivirus:
738 From Empirical Knowledge to Predictive Models. *Journal of Molecular Evolution* **89**,
739 329-340 (2021).
- 740
741
742

743 **Figure Legends**

744

745 **Figure 1:** Rapid expansion of metagenomic-based virome studies and novel viral sequences over
746 time. In green: The number of studies published in NCBI's PubMed database each year from 2001
747 to 2021 that report metagenomic virus discovery/virome analyses [Search query: (metagenomic
748 OR metatranscriptomic) AND (virus OR virome)]. In blue: The number of new virus organisms
749 published in NCBI's nucleotide database each year from 2001 to 2021, sorted by species name.
750 Below the graph, key events in the development of metagenomics are indicated.

751

752 **Figure 2:** Current minimum standards, and how they may be applied to metagenome-assembled
753 viral genomes. Standards outlined in MIMARKS/MiXs are those outlined in Yilmaz *et al.* (2011)²⁶,
754 those from MIUVIG are those outlined in Roux *et al.* (2019)⁸, and those included in a standard
755 virus genome are those outlined in Ladner *et al.* (2014)⁴¹. (Abbreviations -- INSDC: International
756 Nucleotide Sequence Database Collaboration; SRA: Sequence Read Archive; DDBJ: DNA Data
757 Bank of Japan; UViG: Uncultivated Virus Genome; DRA: DDBJ Sequence Read Archive; ENA:
758 European Nucleotide Archive; rRNA: ribosomal RNA; ORF: Open Reading Frame)

759

760 **Figure 3.** Summary of data presentation features we propose for inclusion in all virome studies

761

762 **Figure 4.** Papers published in 2021 using virome scale methods of non-human vertebrate hosts
763 demonstrate many of our recommendations are already being considered by the community. (A)
764 Pie chart of the hosts of virome studies assessed here. (B) Detailed assessment of the 10
765 recommendations proposed here in studies performed in animal hosts. The scoring system
766 included whether each recommendation was (i) fully included as stated here, (ii) partially included
767 such that only some aspects of the recommendation were incorporated, or (iii) whether the
768 recommendation was not considered.

769

770 **Figure 5.** Current state of host association. (A) As indicated in dark orange, a substantial
771 proportion of viral records in GenBank do not have an associated host. Modified from Cobbin *et al.*
772 (2021)²⁷. (B) Eleven picornaviruses recovered from bat faeces, all including "bat" in the virus name,
773 have 3 different hosts as indicated by silhouettes overlaying the plot. Modified from Yinda *et al.*
774 (2017)⁶³.

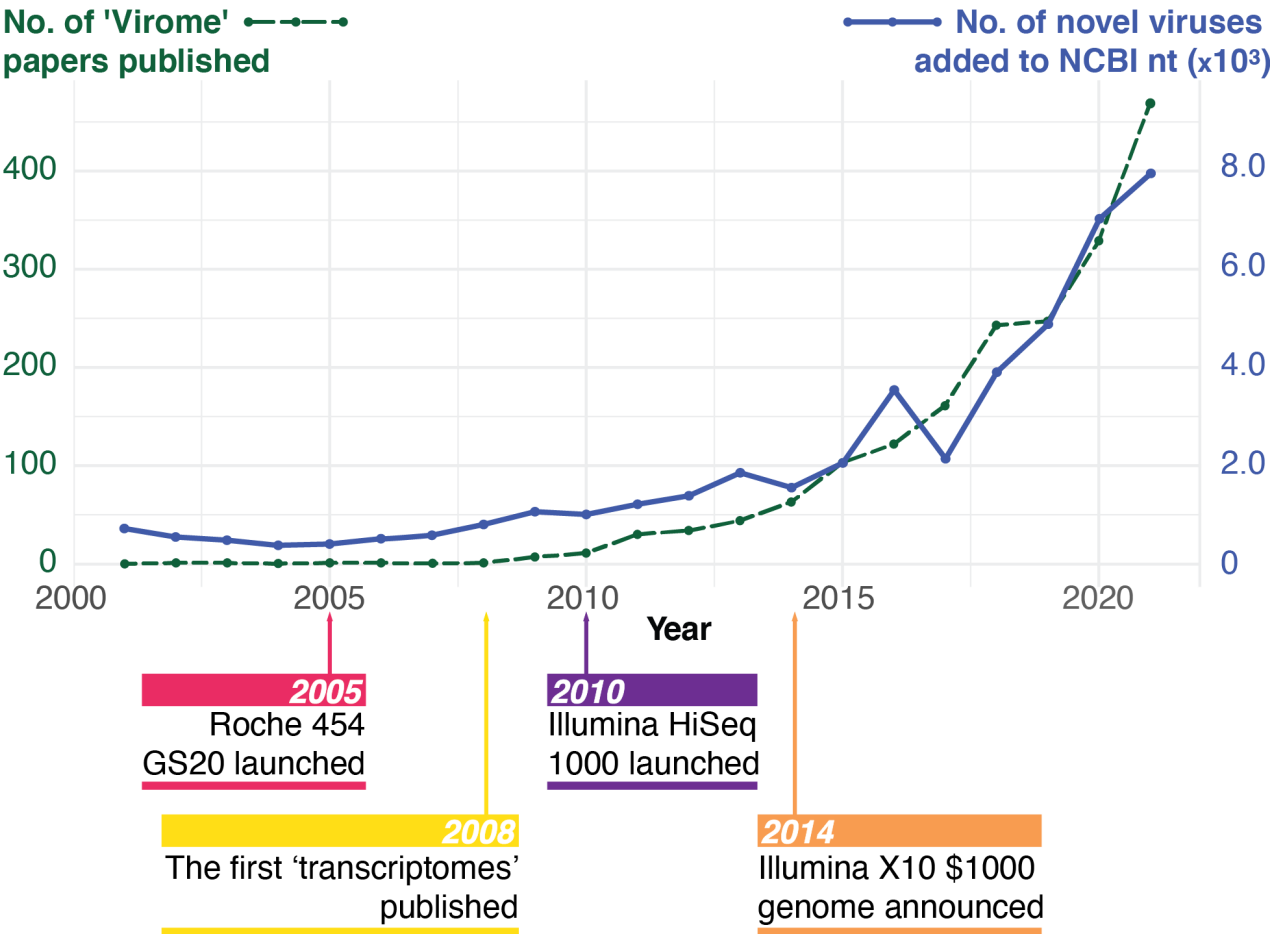
775

776

777 Figure 1.

778

779



780

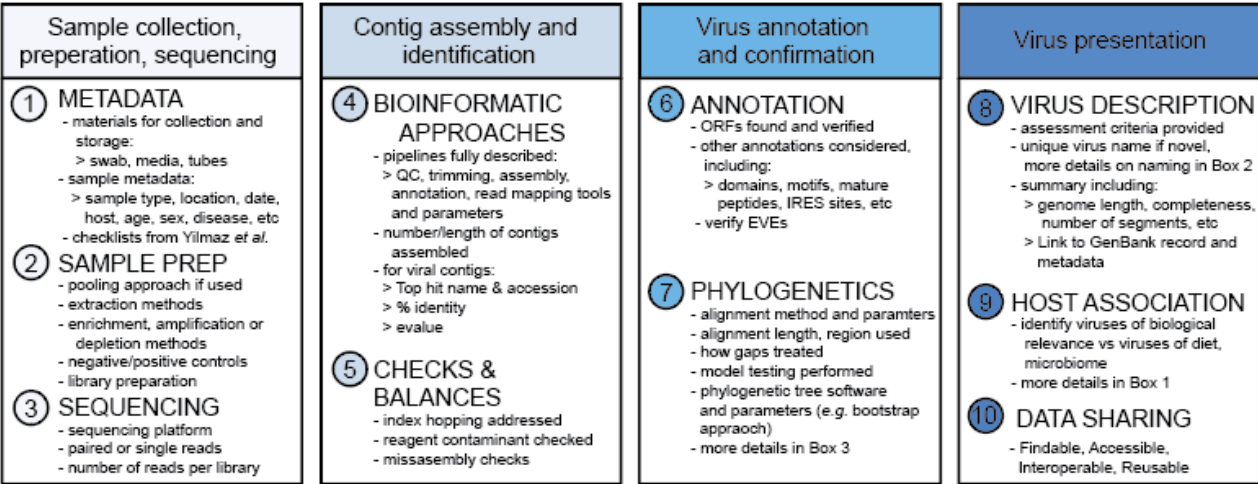
781

782 Figure 2
783

	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

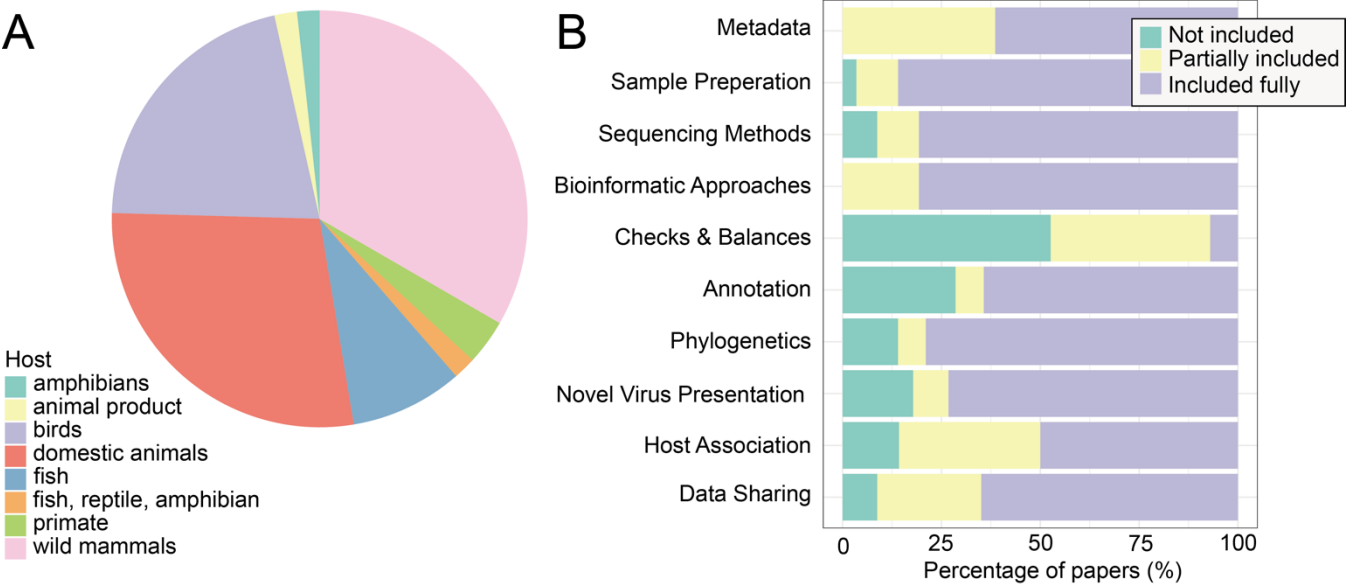
784
785

786 Figure 3
787



788
789

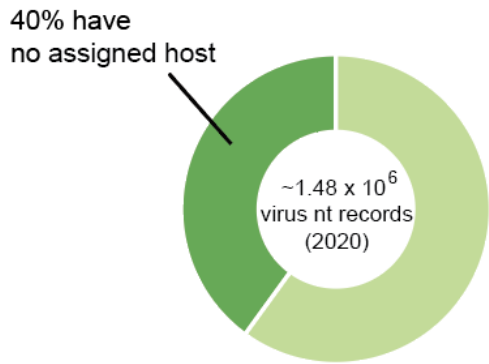
790 Figure 4
791



792
793

794 Figure 5
795

A. Proportion virus nt records
with host association



B. Virus names in public databases
may have misleading names

