Guidelines for public database submission of uncultivated virus genome sequences for taxonomic classification

Evelien M Adriaenssens, Simon Roux, J. Rodney Brister, Ilene Karsch-Mizrachi, Jens H Kuhn, Arvind Varsani, Tong Yigang, Alejandro Reyes, Cédric Lood, Elliot J Lefkowitz, et al.

Adriaenssens EM et al., Nat Biotechnol. 2023; 41(7):898-902. doi: 10.1038/s41587-023-01844-2.

Guidelines for public database submission of uncultivated virus genome sequences for taxonomic classification

Evelien M. Adriaenssens[1*], Simon Roux[2], J. Rodney Brister[3], Ilene Karsch-Mizrachi[3], Jens H. Kuhn[4], Arvind Varsani[5,6], Tong Yigang[7], Alejandro Reyes[8], Cédric Lood[9,10,20], Elliot Lefkowitz[11], Matthew B. Sullivan[12,13,14], Robert A Edwards[15], Peter Simmonds[16], Luisa Rubino[17], Sead Sabanadzovic[18], Mart Krupovic[19], Bas E Dutilh[20,21]

[1] Quadram Institute Bioscience, Norwich Research Park, Rosalind Franklin Road, NR2 7UQ Norwich, United Kingdom *Corresponding author
[2] United States Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[3] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[4] Integrated Research Facility at Fort Detrick, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Fort Detrick, Frederick, MD 21702, USA
[5] The Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University, Tempe, AZ 85287, USA
[6] Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Cape Town 7925, South Africa
[7] Beijing Advanced Innovation Center for Soft Matter Science and Engineering, College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China
[8] Max Planck Tandem Group in Computational Biology, Department of Biological Sciences, Universidad de los Andes, Bogotá, 111711, Colombia.
[9] Centre of Microbial and Plant Genetics, Department of Microbial and Molecular Systems, KU Leuven, 3000 Leuven, Belgium
[10] Laboratory of Gene Technology, Department of Biosystems, KU Leuven, 3000 Leuven, Belgium
[11] Department of Microbiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA
[12] Departments of Microbiology, The Ohio State University, Columbus, OH 43210, USA
[13] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH 43210, USA
[14] Center of Microbiome Science, The Ohio State University, Columbus, OH 43210, USA
[15] College of Science and Engineering, Flinders University, Bedford Park, South Australia 5042, Australia
[16] Nuffield Department of Medicine, University of Oxford, South Parks Road, OX1 3SY Oxford, United Kingdom
[17] Consiglio Nazionale delle Ricerche, Istituto per la Protezione Sostenibile delle Piante, 70126 Bari, Italy
[18] Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Mississippi State, MS 39762, USA
[19] Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, 75015 Paris, France
[20] Institute of Biodiversity, Faculty of Biological Sciences, Cluster of Excellence Balance of the Microverse, Friedrich Schiller University Jena, 07743 Jena, Germany
[21] Theoretical Biology and Bioinformatics, Department of Biology, Science for Life, Utrecht University, 3584 CH Utrecht, Netherlands

Corresponding author: Evelien M. Adriaenssens, evelien.adriaenssens@quadram.ac.uk

To the editor

Mining data derived from high throughput DNA or RNA sequencing approaches, including metagenomics, has led to the discovery of a multitude of uncultivated virus genome sequences [1–12]. These sequences improve our knowledge of the representation of the global virosphere and fuel the expansion and refinement of virus taxonomy. Inclusion of these newly discovered viral sequences into high-quality reference databases is a bottleneck to virology. For formal taxonomic classification, International Committee on Taxonomy of Viruses (ICTV) guidelines stipulate that genome sequences have to be available from a public database. However, the correct use of nomenclature and inclusion of standardized metadata fields is equally as important as the availability of the sequence data to enable the use and reuse of the data by the global research community. Here, we present standards and recommendations for the submission of virus genome sequence data to public databases for the purpose of taxonomic classification. These represent a conceptual and practical extension to the Minimum Information about an Uncultivated Virus Genome (MIUViG) standards that include standards on reporting the virus origin, genome quality, genome annotation, taxonomic classification, biogeographic distribution and host prediction[13]. Aspects of these standards have been reiterated in a recently published consensus view stating that viruses inferred from metagenomic sequences require strict quality control before they can be used for taxonomic assignments[14]. The guidelines presented here focus on the MIUViG standards on genome quality and expand on naming of sequences and submission to public databases.

ICTV coordinates the classification of viruses into 15 taxonomic ranks from species up to realm[15–17] (Figure 1). It is important to note that the ICTV is not responsible for the classification of viruses below the rank of species, such as strains, variants, isolates, lineages, genotypes, or serotypes within individual species, which are instead generally classified by community consensus over time or by non-ICTV expert groups[18,19]. At the species rank, the ICTV requires that the complete genome sequence of a representative member or “exemplar virus” (isolated or identified by [meta]genomic sequencing) is available as an annotated sequence record in one of the International Nucleotide Sequence Database Collaboration (INSDC) member databases [20]. Practically, this means that the annotated genome sequence of any exemplar virus should be submitted to GenBank (National Center for Biotechnology Information [NCBI]), the European Nucleotide Archive (ENA), or the DNA Data Bank of Japan (DDBJ) [21,22]. This choice was guided by the long-term proven reliability, global accessibility, and visibility of INSDC databases. Due to this requirement, at least one fully sequenced virus genome per ICTV-ratified species is now readily available to the global research community and can be used as a reference in comparative genomics analyses.

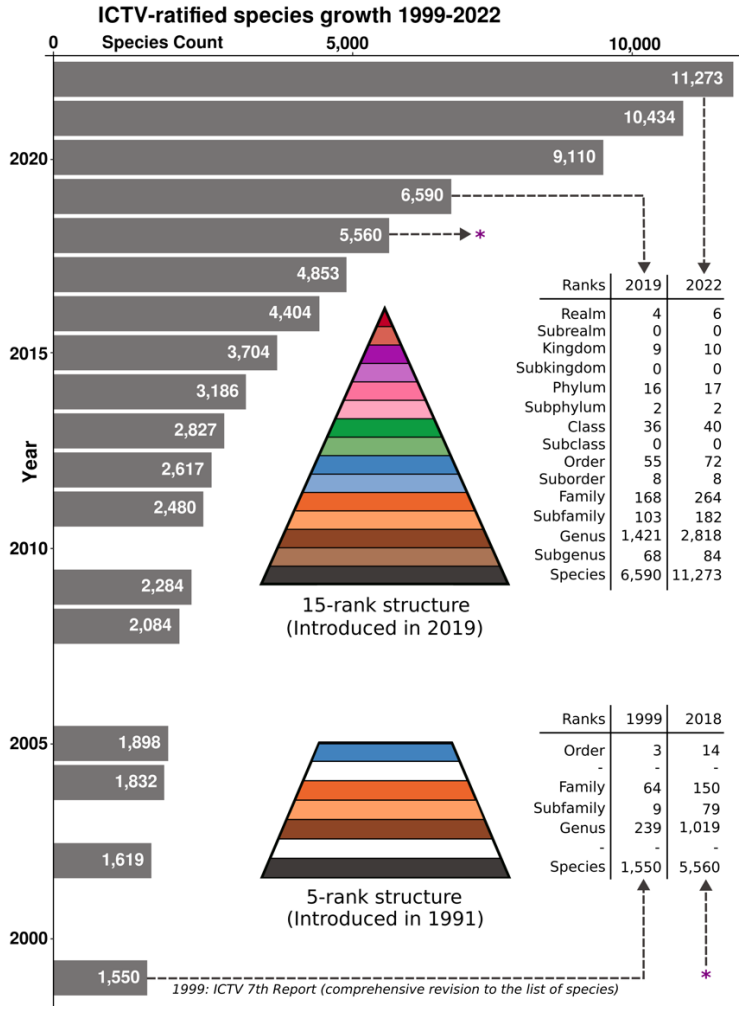

Figure 1: Growth in ICTV-ratified species numbers since the 7th ICTV Report in 1999. The report in 1999 was based on a five-rank structure that was introduced in 1991. The 15-rank taxonomic structure that comprised new ranks such as class, phylum, kingdom, and realm, was introduced in 2019. This figure illustrates the ongoing increase in the number of assigned taxa and the framework that allows classification of UViGs.

We note that many complete, coding-complete, and incomplete virus genome sequences and genomic fragments are available in public repositories other than INSDC (e.g., IMG/VR[12], BV-BRC[23], RAST[24], iVirus[25] or GISAID[26]), whereas other databases such as the Sequence Read Archive (SRA) and Whole Genome Shotgun (WGS) contain unassembled sequencing reads and unannotated or draft genomes, respectively (example guidance from NCBI: https://www.ncbi.nlm.nih.gov/sra/docs/submit/ and https://www.ncbi.nlm.nih.gov/genbank/wgs/). Such repositories provide a resource for data mining of virus genome sequences if these genomes are further assembled and annotated[27,28]. By mandating the deposition of annotated sequences into the INSDC databases, ICTV limits the scattering of exemplar genome sequences across databases and promotes the accessibility of the taxonomically-classified exemplar viruses. Furthermore, the close links between the ICTV and INSDC through NCBI enables better database organization and updating because taxonomy identifiers are persistent and the identifiers are updated routinely with each new ICTV taxonomy release.

A virus genome sequence may be submitted to INSDC databases using the dedicated portals of NCBI (BankIt or table2asn), ENA (Webin), or DDBJ (Nucleotide Sequence Submission System [NSSS]), choosing the submission route for individual complete genomes, or through batch submission. If the virus genome sequence was assembled from datasets that were generated by the submitter, submission follows the same protocols as submission of a virus isolate genome. The sequencing reads should be deposited in the SRA database with the metadata linked through BioProject and BioSample[29], which contain biological data related to individual initiatives (projects) and descriptions of biological source materials (samples) respectively. Metadata in these databases are provided in structured ontologies including the Biological Sample Ontology, the Environment Ontology[30], and the Disease Ontology. Although the availability of raw data cannot be enforced and no mandatory requirements currently exist from the ICTV, submitting such data is a best practice that will be useful for future work, including virus discovery and population genetics studies.

If a genome sequence was assembled from a public dataset, submission to an INSDC database should be done as a Third Party Annotation (TPA), a protocol that was initiated for cases where the original data does not belong to the submitter (see http://www.insdc.org/tpa.html for details and Tisza and Buck (2021)[7] for an example). Even when the original dataset is in the public domain, we recommend that – whenever possible – the submitter of a newly (re-) assembled or (re-) annotated genome sequence contacts the original data depositor(s) to communicate that the data are being reused.

Practical aspects of submission to INSDC databases, with GenBank as an example, are briefly discussed here and published as a detailed standalone guide in Supplementary File 1. Practical guidelines for batch submission of Uncultivated Virus Genome (UViG) sequences are provided in Supplemental File 2.

Genome completeness and sequence quality: To be considered valid for taxonomic classification, genome sequences should be properly assembled. Assembled genome sequences should be checked for terminal redundancy or other evidence of genome termini[31], contigs should be checked for chimerism by evaluating the distribution of mapped reads and read pairs, and partially mapped or unmapped reads remaining in the dataset should be assessed and interpreted. The deposited genomes of exemplar viruses should at least be coding-complete, meaning that all open reading frames (ORFs) in the viral genome are fully sequenced [32], whereas genomic non-coding terminal regions or repeat sequences may be incomplete. Incomplete genome sequences or fragments can still be used to provide context for taxonomic classification, but a coding-complete genome sequence is always required to establish a new taxon. More detailed comments and recommendations on genome sequence completeness can be found in Supplementary File 1, sections 1&3.

UViG sequence submission and naming: GenBank requires every sequence record to have a species-rank taxonomic assignment in the <ORGANISM> field. A problem arises when a sequence belongs to a species that was not previously established. In such cases, a species-rank node is created and named according to the format "<lowest fitting taxon> sp.", in which the <lowest fitting taxon> consists of the formal ICTV name of the lowest ranking taxon that can be confidently assigned according to the demarcation criteria and "sp." for "species" indicates a novel species that has not yet been taxonomically established and named (Figure 2). Examples are "*Sapovirus* sp.", "*Herelleviridae* sp.", and "*Cressdnaviricota* sp.". There is currently no ICTV-approved method to automatically assign a virus query sequence to its lowest fitting taxon because demarcation criteria for assigning sequences to taxa vary widely and should be cross-referenced with taxonomy proposals. Viral ecologists have

defined operational clustering of viral sequences into viral operational taxonomic units (vOTUs) based on universal sequence similarity cutoffs[13], but ICTV-ratified taxa go beyond such preliminary clusters by ensuring some robustness and providing additional information about the members of a taxon. In the GenBank record, metagenomic sequences should be given the /metagenomic, /metagenome_source="…" and /environmental_sample source qualifiers. If further study shows that some or all the sequences in a metagenomic set have been misclassified, submitters may request an update (https://www.ncbi.nlm.nih.gov/genbank/update/) and GenBank will rename and reclassify the sequences, e.g., from "*Siphoviridae* sp." to "*Vequintavirinae* sp.". GenBank may also update the organism name in the record, e.g., from "*Sapovirus* sp." to "*Herelleviridae* sp." without submitter's approval if ICTV sequence analysis indicates that a virus containing an "sp." label has been misfiled.

Using the GenBank record format as a model (Figure 2), we recommend the following:

- <DEFINITION>: This field is automatically populated from the features in the record using a combination of <ORGANISM> and <ISOLATE> name.
- <ORGANISM>: For UViGs, enter the "<lowest fitting taxon> sp.". For an isolate, enter the virus name.
- <ISOLATE>: Enter a unique name/code to describe this specific virus genome sequence. Ensure that this field is unique and is unlikely to be used in another study. Do not use taxonomy information in this field, because virus taxonomy is dynamic. As viruses are reclassified, taxonomy information in the <ORGANISM> field will automatically update, but isolate and genome designations are stable over time and hence should not be at odds with taxonomic names. For example, a novel virus <ISOLATE> should not be called "novel flavivirus 5", as it may turn out not to be a flavivirus in the current or future classification.
- Most databases can, at present, only accommodate the 26 letters of the Medieval Latin alphabet (i.e., ISO basic), ten numbers, and a few special characters, such as hyphens, underscores, and forward slashes. If an official virus name contains Greek letters, special characters or diacritics (e.g., Đakrông virus), feel free to enter them but be aware that most databases will convert them to the standard Latin-script letters (e.g., Dakrong virus), or may even produce an error; the correct spelling in publications should remain Đakrông virus. Underscores and hyphens may be used; forward slashes are typically included in IDs for virus pathogens with formatting requirements, such as members of *Filoviridae* [19], *Caliciviridae*, and influenza A/B/C/D viruses.
- Critical UViG metadata including assembly methods and sequence quality descriptors can be added as structured comments based on the Minimum Information about any (x) Sequence (MIxS) and MIUViG checklists. The most important MIUViG fields are listed in Table 1.
- Do not use a "complete genome" tag for the virus isolate/genome name unless it has been experimentally verified as complete (including termini determination by, for instance, rapid amplification of complementary DNA [cDNA] ends [RACE]). Currently, the only alternative to "complete genome" in GenBank is "partial genome", which should be used in case of UViGs. To specify the genome completeness, we suggest using the categories from the MIUViG checklist as structured comments, with information about the prediction method provided in the genome metadata (Table 1, Supplementary File 1).

Providing appropriate metadata: In INSDC databases, general sequence metadata such as the origin and source of isolation are stored as source modifiers (see more detailed description in Supplementary File 1, section 4). Using the principles of findability, accessibility, interoperability, and reusability (FAIR) for data stewardship [33], all metadata fields should be provided as structured ontology terms (e.g., The Environment Ontology [30], see also Supplementary File 1). The minimum recommended source modifiers to be used are <ISOLATION SOURCE>, <COLLECTION DATE>, and <COUNTRY>, with

 reserved for viruses with segmented genomes. Additional information specific to UViGs should be provided by submitting a MIUViG sequence[13] metadata checklist[34,35] for each UViG sequence and connecting the resulting BioSample package to the UViG genome sequence record by linking the BioSample ID to the GenBank submission. The definition, format, and expected values for each field in the MIUViG sequence checklist are available on the Genomic Standards Consortium (GSC) website. We refer to the GenBank Nucleotide record OP880254 as an example of how to implement the MIUViG standards (https://www.ncbi.nlm.nih.gov/nuccore/OP880254.1).

Features: Sequence annotations, such as ORFs, introns, encoded proteins, and regulatory elements, are stored as features. Feature annotations should be provided for all UViG sequences that are to be used as exemplar genomes to represent new species. At a minimum, the coding sequences should be specified, including functional annotations based on homology searches, phylogenetic analysis, and conserved protein domains, which should be labelled "putative" until experimentally validated.

The availability of complete and consistently annotated records is crucial for the use and reuse of virus sequences and advancing the virology research field. We aim to assist and support the virology community in its expanding use of (meta-) genomic data and the associated taxonomic efforts by promoting the use of this set of standards. While our recommendations are primarily aimed at viruses inferred from metagenome data (UViGs), they are universally applicable to all viruses. Our capacity to generate sequences still outpaces our ability to classify them, so submitting new virus data according to these outlined guidelines will greatly facilitate their findability, accessibility, and reusability as ICTV strives to build a robust virus taxonomy.

Table 1: Information to provide when submitting UViG sequences to INSDC databases.

Information to provide	Where to add	Description	Suggested syntax[a]
organism	Submission portal + MIUViG checklist structured comment	UViG: lowest ranking taxon that can be confidently assigned according to ICTV demarcation criteria. Isolated virus: virus name.	[<"lowest fitting taxon" sp.> | virus name]
isolate	Submission portal + MIUViG checklist structured comment	Unique name or code for this sequence. Do not use taxonomic information here.	<Unique identifier>
Source of UViG	MIUViG checklist structured comment	Type of sample used for UViG assembly	[metagenome (not viral targeted) | viral fraction metagenome (virome) | sequence-targeted metagenome | metatranscriptome (not viral targeted) | viral fraction RNA metagenome (RNA virome) | sequence-targeted RNA metagenome | microbial single

			amplified genome (SAG) | viral single amplified genome (vSAG) | isolate microbial genome | other]
Assembly software	MIUViG checklist structured comment	Tool(s) used for assembly and optionally binning. Include version and parameters.	{software};{version};{parameters}
Assembly quality	MIUViG checklist structured comment	Assembly quality in categories as per the MIUViG criteria. Finished: Single, validated, contiguous sequence per replicon without gaps or ambiguities, with extensive manual review and annotation. High-quality draft genome: One or multiple fragments, totalling ≥ 90% of the expected genome or replicon sequence or predicted complete. Genome fragment(s): One or multiple fragments, totalling < 90% of the expected genome or replicon sequence, or for which no genome length could be estimated.	[Finished genome | High-quality draft genome | Genome fragment(s)]
Completeness score	MIUViG checklist structured comment	(Optional) Estimated completeness of the UViG in percentage.	{quality};{percentage}
Completeness approach	MIUViG checklist structured comment	(Optional) Approach used to estimate completeness, such as identification of terminal repeats or presence of all CDS	{text}
Virus identification software	MIUViG checklist structured comment	Tool(s) used for identification of sequence as virus. Include versions and parameters.	{software};{version};{parameters}

Predicted genome type	MIUViG checklist structured comment	Type of genome predicted for the UViG.	[DNA | dsDNA | ssDNA | RNA | dsRNA | ssRNA | ssRNA (+) | ssRNA (-) | mixed | uncharacterized]

[a] entries between []: choose one of the listed descriptors; entries between <>: fill in the UViG or virus information for this record; entries between {}: enter data for your methods used.

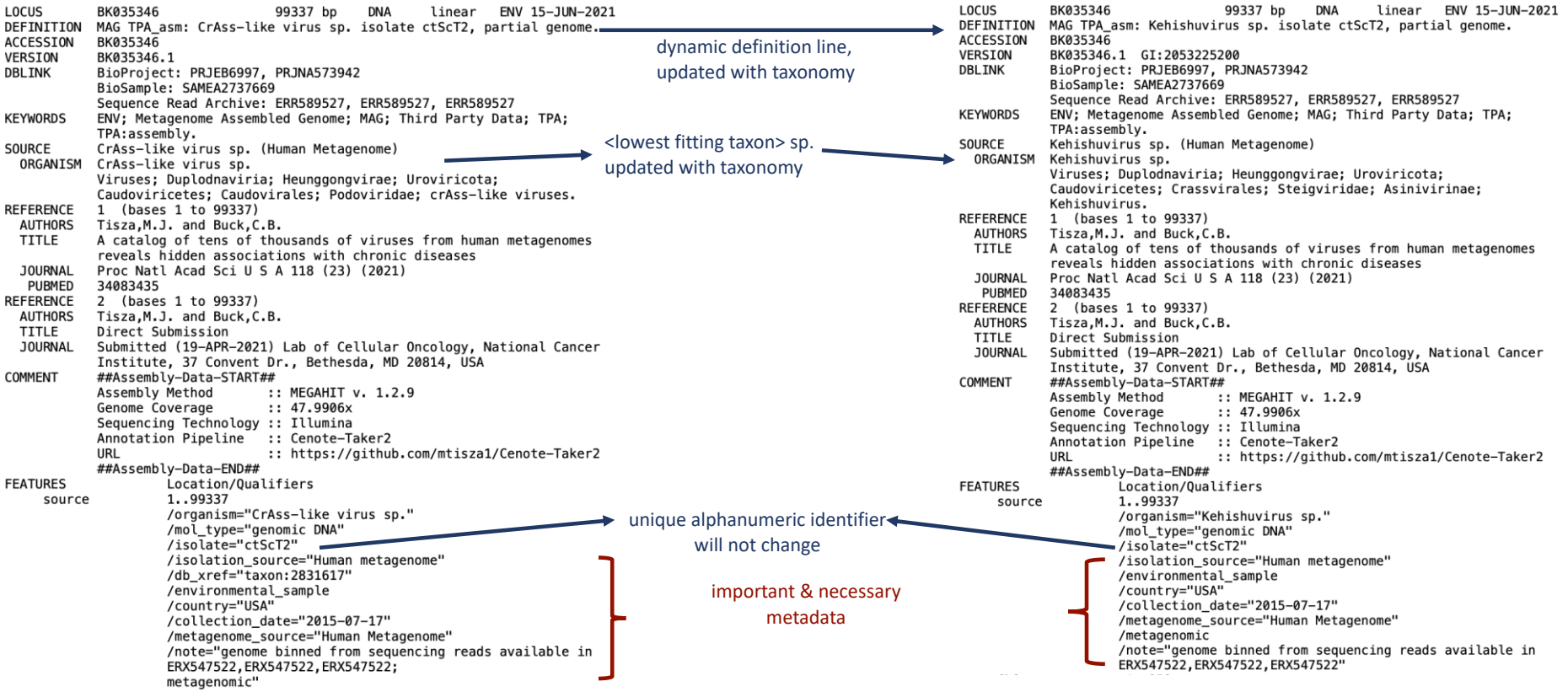

Figure 2: GenBank example of record BK035346. Left: as submitted with the taxonomy at the time of submission; Right: updated GenBank record after a later update to the International Committee on Taxonomy of Viruses (ICTV) taxonomy. The ORGANISM name was updated from CrAss-like virus sp. to *Kehishuvirus* sp. now showing the new taxonomic lineage information. The DEFINITION line was updated according to the ORGANISM change.

Acknowledgements

The authors thank Anya Crane (Integrated Research Facility at Fort Detrick/National Institute of Allergy and Infectious Diseases/National Institutes of Health, Fort Detrick, Frederick, MD, USA) for critically editing the manuscript. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Health and Human Services or of the institutions and companies affiliated with the authors, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

E.M.A. gratefully acknowledges the support of the Biotechnology and Biological Sciences Research Council (BBSRC); this research was funded by the BBSRC Institute Strategic Program Gut Microbes and Health BB/R012490/1 and its constituent project(s) BBS/E/F/000PR10353 and BBS/E/F/000PR10356. The work conducted by the U.S. Department of Energy Joint Genome Institute (S.R.) was supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Work by J.R.B. and I.K.M. was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health. This work was supported in part through Laulima Government Solutions, LLC, prime contract with the U.S. National Institute of Allergy and Infectious Diseases (NIAID) under Contract No. HHSN272201800013C. J.H.K. performed this work as an employee of Tunnell Government Services (TGS), a subcontractor of Laulima Government Solutions, LLC, under Contract No. HHSN272201800013C. MBS was supported by the US National Science Foundation Award #1759874. R.A.E. was supported by the National Institute of Diabetes And Digestive and Kidney Diseases of the National Institutes of Health under Award Number RC2DK116713 and by the Australian Research Council under Award Number DP220102915. C.L. was supported by a Postdoctoral Mandate from KU Leuven (PDMt2/21/038) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2051 – Project-ID 390713860. E.J.L. was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number U24AI162625. P.S. was supported by a Wellcome Trust Biomedical Resource grant (WT108418AIA). S.S. acknowledges support from the Mississippi Agricultural and Forestry Experiment Station (MAFES), USDA-ARS project 58-6066-9-033 and the National Institute of Food and Agriculture, U.S. Department of Agriculture, Hatch Project, under Accession Number 1021494. B.E.D. was supported by the European Research Council (ERC) Consolidator Grant 865694: DiversiPHI, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2051 – Project-ID 390713860, the Alexander von Humboldt Foundation in the context of an Alexander von Humboldt-Professorship founded by German Federal Ministry of Education and Research, and the European Union's Horizon 2020 research and innovation program, under the Marie Skłodowska-Curie Actions Innovative Training Networks grant agreement no. 955974 (VIROINF).

The authors do not declare any conflicts of interest.

References

1. Callanan, J. et al. Expansion of known ssRNA phage genomes: From tens to over a thousand. *Sci. Adv.* 6, eaay5981 (2020).

2. Gregory, A. C. et al. The Gut Virome Database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* 28, 724-740.e8 (2020).

3. Zayed, A. A. et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* 376, 156–162 (2022).

4. Hillary, L. S., Adriaenssens, E. M., Jones, D. L. & McDonald, J. E. RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels. *ISME Commun.* 2, 34 (2022).

5. Roux, S., Krupovic, M., Poulet, A., Debroas, D. & Enault, F. Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One* 7, e40418 (2012).

6. Krishnamurthy, S. R., Janowski, A. B., Zhao, G., Barouch, D. & Wang, D. Hyperexpansion of RNA Bacteriophage Diversity. *PLoS Biol.* 14, e1002409 (2016).

7. Tisza, M. J. & Buck, C. B. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proc. Natl. Acad. Sci.* 118, e2023202118 (2021).

8. Gregory, A. C. et al. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* 177, 1109-1123.e14 (2019).

9. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* 184, 1098-1109.e9 (2021).

10. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* 6, 960–970 (2021).

11. Emerson, J. B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* 3, 870–880 (2018).

12. Roux, S. et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* 49, D764–D775 (2021).

13. Roux, S. et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* 37, 29–37 (2019).

14. Simmonds, P. et al. Four principles to establish a universal virus taxonomy. *PLoS Biol.* 21, e3001922 (2023).

15. Koonin, E. V. et al. Global organization and proposed megataxonomy of the virus world. *Microbiol. Mol. Biol. Rev.* 84, e00061-19 (2020).

16. Gorbalenya, A. E. et al. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.* 5, 668–674 (2020).

17. Simmonds, P. et al. Consensus statement: Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* 15, 161–168 (2017).

18. Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 5, 1403–1407 (2020).

19. Kuhn, J. H. et al. Virus nomenclature below the species level: A standardized nomenclature for natural variants of viruses assigned to the family *Filoviridae*. *Arch. Virol.* 158, 301–311 (2013).

20. Karsch-Mizrachi, I., Nakamura, Y. & Cochrane, G. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 40, D33–D37 (2012).

21. Salzberg, S. L. Reminder to deposit DNA sequences. *Nature* 533, 179–179 (2016).

22. Blaxter, M. et al. Reminder to deposit DNA sequences. *Science* 352, 780–780 (2016).

23. Olson, R. D. et al. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res.* 51, D678–D689 (2023).

24. Aziz, R. K. et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9, 75 (2008).

25. Bolduc, B. et al. iVirus 2.0: Cyberinfrastructure-supported tools and data to power DNA virus ecology. *ISME Commun.* 1, 77 (2021).

26. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* 22, (2017).

27. Mokili, J. L., Rohwer, F. & Dutilh, B. E. Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–77 (2012).

28. Paez-Espino, D. et al. Uncovering Earth's virome. *Nature* 536, 425–430 (2016).

29. Barrett, T. et al. BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. *Nucleic Acids Res.* 40, 57–63 (2012).

30. Buttigieg, P. L. et al. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semantics* 7, 57 (2016).

31. Garneau, J. R. et al. High-throughput identification of viral termini and packaging mechanisms in virome datasets using PhageTermVirome. *Sci. Rep.* 11, 18319 (2021).

32. Ladner, J. T. et al. Standards for sequencing viral genomes in the era of high-throughput sequencing. *MBio* 5, e01360-14-e01360-14 (2014).

33. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018 (2016).

34. NCBI. BioSample Types and Attributes checklists. Available at: https://submit.ncbi.nlm.nih.gov/biosample/template/. (Accessed: 24th May 2023)

35. Genomics Standards Consortium MIUVIG checklist. Available at: https://genomicsstandardsconsortium.github.io/mixs/MIUVIG/. (Accessed: 24th May 2023)

Supplementary Notes

Supplementary Note 1: Standalone guide for the submission of UViG sequences to the INSDC database

In this guide document, we provide submission examples for GenBank (NCBI)[1]. Submission to the DDBJ and ENA may have slightly different requirements and formats. Please note that data submitted to any of the three resources will be available in all of them, since data is mirrored between the INSDC databases.

1. Genome sequence quality

In 2019, a consensus statement on the Minimum Information criteria for Uncultivated Virus Genome sequences (MIUViG) was published, defining three classes of quality for Uncultivated Virus Genome (UViG) sequences: genome sequence fragments (estimated to be <90% complete), high-quality draft genome sequences (estimated to be ≥90% complete), and complete genome sequences with extensive annotations [2]. The authors of the statement (including several of the authors in this statement) recommended that only complete or coding complete genome sequences can be used as reference (exemplar) genome sequences to establish new species. Genome completeness may be inferred from genomic comparison to related viruses, if the candidate genome can be robustly placed within a cluster of viruses with a well-defined gene content, and/or from the topology of the genome sequence itself, e.g., the detection of direct or inverted terminal repeats. However, estimation of completeness and recovery of complete genome sequences is easier for viruses with circular or circularly permuted genomes than for viruses that have segmented/multipartite genomes, or linear genomes with defined termini. Important to note, virus sequences belonging to all three UviG quality categories may be used to provide additional information for the establishment of new taxa, for example, to test the robustness of phylogenetic trees. Complete or coding-complete genome sequences are necessary, however, to serve as exemplars for the establishment of new species.

2. UViG sequence submission and naming

ICTV is concerned with the naming of virus taxa ranging from species to realms [3], but the naming of individual viruses is outside the ICTV responsibility [4]. Here, we provide a set of recommendations and best practices for the labeling of UViG sequences and submission of metadata.

Submitters should provide unique identifiers (IDs) for each sequence in the <ISOLATE> field, preferably as a single string of at least six alphanumerical characters (e.g., blue53F), using hyphens and underscores to tie separate elements together, e.g., “0815_Eier-kuchen”. Submitters should avoid including common terms like “scaffold” or “contig” in the isolate IDs, or IDs that may be used in other studies (e.g., “soil_virus_contig_01” or “phage_P1”).

Sequences from metagenomic sets should be submitted to GenBank in the <ORGANISM> name format “<lowest fitting taxon> sp.”, in which the <lowest fitting taxon> consists of the formal ICTV taxon name rank (genus or higher) that can be confidently assigned to the sequence, by using the

demarcation criteria for each of these ranks (Figure 1). Examples are "*Sapovirus* sp.", "*Herelleviridae* sp.", and "*Cressdnaviricota* sp." [note that in writing taxon names need to be italicized, but italics are not supported by INSDC databases]. Unique <organism name>s for metagenomic sequences, e.g., "*Sapovirus* sp. Seal/X17", are still acceptable if those <organism name>s have been used in publications, e.g., for viruses of medical importance. GenBank will place these "Taxon sp." Names into unclassified bins reserved for non-ICTV names, e.g., "*Sapovirus* sp." Is found within "unclassified Sapovirus", "*Herelleviridae* sp." In "unclassified *Herelleviridae*", and "*Cressdnaviricota* sp." In "unclassified *Cressdnaviricota*".

In the GenBank record, metagenomic sequences should be given the /metagenomic, /metagenome_source="…" and /environmental_sample source qualifiers. If further study shows that some or all the sequences in a metagenomic set have been misclassified, submitters may request an update (https://www.ncbi.nlm.nih.gov/genbank/update/) and GenBank will rename and reclassify the sequences, e.g., from "*Siphoviridae* sp." To "*Vequintavirinae* sp.". INSDC may also update the organism name in the record, e.g., from "*Sapovirus* sp." To "*Herelleviridae* sp." Without submitter approval if ICTV sequence analysis indicates that a virus containing an "sp." Label has been misfiled.

If a sequence originally submitted with a metagenome name, such as, "*Herelleviridae* sp." Is later used as an ICTV exemplar, the INSDC Taxonomy group at NCBI will rename the <organism name> in the sequence record without requiring submitter approval upon processing the release of the new taxonomy. This information is stored and communicated through the Virus Metadata Resource (VMR, the ICTV file linking the taxonomy with the GenBank accession numbers, https://talk.ictvonline.org/taxonomy/vmr/).

3. Submission recommendations for naming and completeness

In summary, using the GenBank record format as a model (Figure 1), we recommend the following:

- <DEFINITION>: This field is automatically populated from the features in the record using a combination of <ORGANISM> and <ISOLATE> name.
- <ORGANISM>: Enter "<lowest fitting taxon> sp.".
- <ISOLATE>: Enter a unique name/code to describe this specific virus genome sequence. Ensure that this field is unique and is unlikely to be used in another study. Do not use taxonomy information in this field, because virus taxonomy is dynamic. As viruses are reclassified, taxonomy information in the <ORGANISM> field will automatically update, but isolate and genome designations are stable over time and hence should not be at odds with taxonomic names. For example, a novel virus <ISOLATE> should not be called "novel flavivirus 5", as it may turn out not to be a flavivirus in the current or future classification.
- Names should take into account that most databases can, at present, only accommodate the 26 letters of the Medieval (aka ISO basic) Latin alphabet, numbers, and a few special characters, such as, hyphens. If a virus name contains Greek letters, special characters or diacritics (e.g., Đakrông virus), feel free to enter them but be aware that most databases will convert them to the standard Latin-script letters (e.g., Dakrong virus) or produce an error; the correct spelling in publications will remain Đakrông virus. Underscores and hyphens can be used; forward slashes are typically included in IDs for virus pathogens with formatting requirements, such as, members of *Filoviridae* [5], *Caliciviridae*, and influenza viruses.
- Do not use a "complete genome" tag for the virus isolate/genome name unless it has been experimentally verified as complete (including termini determination by, for instance, rapid amplification of complementary DNA [cDNA] ends [RACE]). Genomes that have been

bioinformatically predicted as being complete may be identified as “predicted complete genome”, with information about the prediction method provided in the genome metadata. Note that, in GenBank, the only alternative to “complete” is “partial”, and as a result, the vast majority of UViGs will be tagged as partial genomes. It is the authors’ opinion, that this strict criterion could be reassessed as more computational methods are validated and that the specific MIUViG completeness scores added as structured comments could be used in future to provide more nuance. In GenBank, viral genomes will not be labelled complete if they contain a stretch of 100 or more ambiguous characters.

- Submit genome metadata via the “Source Modifiers” section of the genome submission process (for general metadata). Additionally, the creation of a separate BioSample for each genome sequence is encouraged using the Minimum Information about any (x) Sequence (MixS) “MIUViG” checklist (for UViG-specific metadata). The metadata fields for UViG quality and completeness (see also Table 1) should be added as structured comments.

4. Providing appropriate metadata

Source modifiers

In INSDC databases, metadata information on a sequence is stored in source modifiers. Using the principles of findability, accessibility, interoperability, and reusability (FAIR) for data stewardship [6], it is best practice to provide as much source metadata as possible, by using structured ontology terms (e.g., The Environment Ontology [7]). Here, we offer guidelines on the implementation of commonly used source modifiers that may be used to provide structured metadata information.

- <HOST> field: Use this field for the host from which the sample was isolated. We recommend not using this source modifier and instead using the MIUViG checklist (see below) for host prediction or the “isolation source” field to provide sample-specific information. If the virus host is predicted from the sequence using computational means [8], the confidence score should be reported (expected precision). Otherwise, leave this field blank. Use the taxonomy IDs from NCBI taxonomy for host description.
- <ISOLATION SOURCE>: Use this field to describe the sample from which the sequence was derived using the Environment Ontology[7] (see also https://www.ebi.ac.uk/ols/index).
- <COLLECTION DATE>: Enter the date of collection for the sample from which the sequence was obtained in the format YYYY-MM-DD.
- <COUNTRY>: Enter the country in which the sample was collected. A standardized list of countries for INSDC submissions can be found here: https://www.ncbi.nlm.nih.gov/genbank/collab/country/.
- : For viruses with segmented genomes, this modifier can be used to indicate which segment was recovered. Use this field only if the genomes are similar enough to those of known viruses, i.e., fall within the published demarcation criteria for inclusion into established species for positive segment identification.
- <NOTES>: Note that free text is difficult to computationally parse and is thus not FAIR compliant. Any information that can be entered using structured ontologies as in the fields above is preferred. Use this free text box to add any information that cannot be accounted for in specific source modifiers.

Features

Sequence annotations, such as ORFs, introns, encoded proteins, and regulatory elements, are stored as features in INSDC. Feature annotations should be provided for all UviG sequences that are to be used as exemplar genome sequences to represent new species. At a minimum, the coding

sequences should be provided, including putative functional annotations based on homology searches, phylogenetic analysis, and conserved protein domains. It is good practice to add as many features as can be identified (e.g., transfer RNA [tRNA], terminal repeat regions, promoters).

UViG sequence-specific metadata for BioSample submission

Most often, UViG sequences are accompanied by specific methodological metadata, including the assembly pipeline, viral sequence identification method, completeness estimation, and host prediction. It is critical to attach this information to a UViG genome sequence record, but it does not fit in the standard set of "source" metadata. Moreover, this information is often predicted by bioinformatic programs and thus remains tentative. Instead, this information should be provided by submitting a MIUViG sequence [2] metadata checklist (https://gensc.org/mixs/submit-mixs-metadata/) for each UViG sequence and connecting the resulting BioSample package to the UViG genome sequence record by linking the BioSample ID to the GenBank submission. The definition, format, and expected values for each field in the MIUViG sequence checklist are available on the Genomic Standards Consortium (GSC) website (https://gensc.org/mixs/), with the most important and mandatory ones being:

- <source_uvig>: Type of dataset from which the UviG sequence was obtained, to be selected from "metagenome (not viral targeted)", "viral fraction metagenome (virome)", "sequence-targeted metagenome", "metatranscriptome (not viral targeted)", "viral fraction RNA metagenome (RNA virome)", "sequence-targeted RNA metagenome", "microbial single amplified genome (SAG)", "viral single amplified genome (vSAG)", "isolate microbial genome", and "other"
- <vir_ident_software>: Tool(s) used for the identification of a UviG sequence as a viral genome, such as, the software or protocol name including version number and the used parameters and cutoffs
- <pred_genome_type>: Type of genome predicted for the UviG sequence, to be selected from "DNA", "dsDNA", "ssDNA", "RNA", "dsRNA", "ssRNA", "ssRNA (+)", "ssRNA (-)", "mixed", and "uncharacterized"
- <pred_genome_struc>: Expected structure of the viral genome, to be selected from "segmented", "non-segmented", and "undetermined"
- <detec_type>: Type of UviG detected to be selected from "independent sequence (UviG)" (separate contig in dataset), "provirus (UpViG)" (sequenced flanked by host DNA)
- <host_pred_appr> and <host_pred_est_acc>: Tool or approach used for host prediction, and estimated false discovery rates for these tools either computed de novo or from the literature

Supplementary Note 2: Practical guidelines for Batch submission of Uncultivated Virus Genome (UViG) sequences to GenBank

The command-line program `table2asn` allows the quick generation of .asn files for submission to GenBank for thousands of sequences at a time. These .asn files can then be uploaded through BankIt. For submissions of more than 5,000 viruses, submitters are encouraged to contact the database administrators to ensure a smooth submission process.

Template: To run table2asn you will first need to generate a template file (https://submit.ncbi.nlm.nih.gov/genbank/template/submission/). These templates may include the BioProject and BioSample accessions, as well as publication information. This template may then be used when running `table2asn` using `-t` in the command line.

Assembly Information: Assembly information can be incorporated into a structured format that can be added while running `table2asn`. To include the assembly data in the .sqn file, create a tab-delimited table in this format:

```
StructuredCommentPrefix          ## Assembly-Data-START##
Assembly Method                  Unicycler v. 0.4
Genome Coverage                  177x
Sequencing Technology            Illumina; Nanopore
StructuredCommentSuffix          ##Assembly-Data-END##
```

Note that the assembly method script requires “v. ” between the algorithm name and its version. If more than one sequencing technology was used, enter both and separate them with a semi-colon.

To include this information when running `table2asn`, use `-w stru_cmt_file` (for which “str_cmt_file” is the name of your assembly information).

Source Information: All source information should be complete and include the mandatory fields isolation_source, collection_date, and country, using the relevant ontologies. Each virus sample must be associated with a unique identifier (as described in this publication) that can be used to separate this virus from other submissions in the database. Source information for viruses inferred from metagenomic data must also contain environmental sample and metagenomic flags and should indicate the type of metagenome from which the samples were obtained (metagenome_source: e.g., fungus metagenome, plant metagenome, gut metagenome).

Source information can be incorporated into the file in multiple ways:

[1] Source information and molecule information can be included in the fasta header of each sequence. For example:
> rainbowtrout_1ct44 [organism=Circovirus sp.] [isolate=ct44] [isolation-source=subsurface seawater] [country=USA] [collection-date=04-Jun-2018] [topology=circular] [BioProject=PRJNAXXXXXX] [BioSample=SAMNXXXXXXXXXX]
Note: If the BioProject and BioSample information were included in the template file, there is no need to include them here.

[2] Some of the source information that is shared by all submissions can be incorporated into the file using the `-j` command in the `table2asn` command line, while source qualifiers unique to each genome can be incorporated using the fasta definition line.
For example:

`-j [isolation-source=subsurface seawater] [country=USA] [collection-date=04-Jun-2018]`

[3] A tab-delimited source modifier table can be created to be read by `table2asn`. The instructions for this table construction can be found at https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html. Use the file suffix .src and match the prefix to the other files.

Feature Annotation: Genomic feature annotation is recommended, but not required for virus submissions, unless they are being used as International Committee on Taxonomy of Viruses (ICTV) species exemplars. Feature annotation can be incorporated using `table2asn` by including `-f feature.tbl` for which feature.tbl is the name of your table. For this to work, the header in each feature table must match the seqID in the corresponding fasta file; i.e., if the fasta file header is ">abcd1", the corresponding feature table should begin with ">Feature abcd1".

Supplementary Note references

1. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2013; 41:36–42.

2. Roux S, Adriaenssens EM, Dutilh BE, Koonin E V., Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). Nat Biotechnol [Internet] 2019; 37:29–37. Available from: http://www.nature.com/doifinder/10.1038/nbt.4306

3. Gorbalenya AE, Krupovic M, Mushegian A, Kropinski AM, Siddell SG, Varsani A, Adams MJ, Davison AJ, Dutilh BE, Harrach B, et al. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. Nat Microbiol 2020; 5:668–74.

4. Kuhn JH, Jahrling PB. Clarification and guidance on the proper usage of virus and virus species names. Arch Virol 2010; 155:445–53.

5. Kuhn JH, Bao Y, Bavari S, Becker S, Bradfute S, Brister JR, Bukreyev AA, Chandran K, Davey RA, Dolnik O, et al. Virus nomenclature below the species level: A standardized nomenclature for natural variants of viruses assigned to the family Filoviridae. Arch Virol 2013; 158:301–11.

6. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data [Internet] 2016; 3:160018. Available from: http://www.nature.com/articles/sdata201618

7. Buttigieg PL, Pafilis E, Lewis SE, Schildhauer MP, Walls RL, Mungall CJ. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. J Biomed Semantics [Internet] 2016; 7:57. Available from: https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-016-0097-6

8. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol Rev [Internet] 2016; 40:258–72. Available from: https://academic.oup.com/femsre/article-lookup/doi/10.1093/femsre/fuv048