

Four principles to establish a universal virus taxonomy

Peter Simmonds, Evelien M Adriaenssens, F. Murilo Zerbini, Nicola G A Abrescia, Pakorn Aiewsakun, Poliane Alfenas-Zerbini, Yiming Bao, Jakub Barylski, Christian Drosten, Siobain Duffy, et al.

▶ **To cite this version:**

Peter Simmonds, Evelien M Adriaenssens, F. Murilo Zerbini, Nicola G A Abrescia, Pakorn Aiewsakun, et al.. Four principles to establish a universal virus taxonomy. PLoS Biology, 2023, 21 (2), pp.e3001922. 10.1371/journal.pbio.3001922 . pasteur-03988048

HAL Id: pasteur-03988048
https://pasteur.hal.science/pasteur-03988048

Submitted on 14 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC0 - Public Domain Dedication 4.0 International License

PLOS BIOLOGY

CONSENSUS VIEW

Four principles to establish a universal virus taxonomy

Peter Simmonds[1]*, Evelien M. Adriaenssens[2]*, F. Murilo Zerbini[3]*, Nicola G. A. Abrescia[4,5], Pakorn Aiewsakun[6], Poliane Alfenas-Zerbini[7], Yiming Bao[8,9], Jakub Barylski[10], Christian Drosten[11,12], Siobain Duffy[13], W. Paul Duprex[14], Bas E. Dutilh[15,16], Santiago F. Elena[17,18], Maria Laura García[19], Sandra Junglen[11,12], Aris Katzourakis[20], Eugene V. Koonin[21], Mart Krupovic[22], Jens H. Kuhn[23], Amy J. Lambert[24], Elliot J. Lefkowitz[25], Małgorzata Łobocka[26], Cédric Lood[27], Jennifer Mahony[28], Jan P. Meier-Kolthoff[29], Arcady R. Mushegian[30], Hanna M. Oksanen[31], Minna M. Poranen[31], Alejandro Reyes-Muñoz[32], David L. Robertson[33], Simon Roux[34], Luisa Rubino[35], Sead Sabanadzovic[36], Stuart Siddell[37], Tim Skern[38], Donald B. Smith[1], Matthew B. Sullivan[39], Nobuhiro Suzuki[40], Dann Turner[41], Koenraad Van Doorslaer[42], Anne-Mieke Vandamme[43,44], Arvind Varsani[45], Nikos Vasilakis[46]

1 Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom, **2** Quadram Institute Bioscience, Norwich Research Park, Norwich, United Kingdom, **3** Departamento de Fitopatologia/BIOAGRO, Universidade Federal de Viçosa, Viçosa, Brazil, **4** Structure and Cell Biology of Viruses Lab, Center for Cooperative Research in Biosciences—BRTA, Derio, Spain, **5** Basque Foundation for Science, IKERBASQUE, Bilbao, Spain, **6** Department of Microbiology, Faculty of Science, Mahidol University, Bangkok, Thailand, **7** Departamento de Microbiologia/BIOAGRO, Universidade Federal de Viçosa, Viçosa, Brazil, **8** National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing, China, **9** University of Chinese Academy of Sciences, Beijing, China, **10** Department of Molecular Virology, Adam Mickiewicz University, Poznan, Poland, **11** Institute of Virology, Charité-Universitätsmedizin Berlin, corporate member of Free University Berlin, Humboldt University, Berlin, Germany, **12** Berlin Institute of Health, Berlin, Germany, **13** Department of Ecology, Evolution and Natural Resources, School of Environmental and Biological Sciences, Rutgers The State University of New Jersey, New Brunswick, New Jersey, United States of America, **14** The Center for Vaccine Research, University of Pittsburgh School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, **15** Institute of Biodiversity, Faculty of Biological Sciences, Cluster of Excellence Balance of the Microverse, Friedrich-Schiller-University, Jena, Germany, **16** Theoretical Biology and Bioinformatics, Science for Life, Utrecht University, Utrecht, the Netherlands, **17** Instituto de Biología Integrativa de Sistemas (I2SysBio), CSIC-Universitat de València, Valencia, Spain, **18** Santa Fe Institute, Santa Fe, New Mexico, United States of America, **19** Instituto de Biotecnología y Biología Molecular, CCT-La Plata, CONICET, UNLP, La Plata, Argentina, **20** Department of Biology, University of Oxford, Oxford, United Kingdom, **21** National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, **22** Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, Paris, France, **23** Integrated Research Facility at Fort Detrick (IRF-Frederick), National Institute of Allergy and Infectious Diseases, National Institutes of Health, Fort Detrick, Frederick, Maryland, United States of America, **24** Division of Vector-Borne Diseases, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Fort Collins, Colorado, United States of America, **25** Department of Microbiology, University of Alabama at Birmingham, Birmingham, Alabama, United States of America, **26** Institute of Biochemistry and Biophysics of the Polish Academy of Sciences, Warsaw, Poland, **27** Department of Biosystems, KU Leuven, Leuven, Belgium, **28** School of Microbiology and APC Microbiome Ireland, University College Cork, Cork, Ireland, **29** Department of Bioinformatics and Databases, Leibniz Institute DSMZ—German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany, **30** Division of Molecular and Cellular Biosciences, National Science Foundation, Alexandria, Virginia, United States of America, **31** Molecular and Integrative Biosciences Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland, **32** Max Planck Tandem Group in Computational Biology, Departamento de Ciencias Biológicas, Universidad de los Andes, Bogotá, Colombia, **33** MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom, **34** Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, **35** Istituto per la Protezione Sostenibile delle Piante, CNR, UOS Bari, Bari, Italy, **36** Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Mississippi State, Mississippi, United States of America, **37** School of Cellular and Molecular Medicine, Faculty of Life Sciences, University of Bristol, Bristol,

OPEN ACCESS

Citation: Simmonds P, Adriaenssens EM, Zerbini FM, Abrescia NGA, Aiewsakun P, Alfenas-Zerbini P, et al. (2023) Four principles to establish a universal virus taxonomy. PLoS Biol 21(2): e3001922. https://doi.org/10.1371/journal.pbio.3001922

Published: February 13, 2023

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Funding: The workshop and D.B.S. were funded by a Wellcome Biomedical Resource grant to P.S. and S.S. (WT108418AIA). M.M.P. acknowledges funding from the Academy of Finland (grant 331627) and the Sigrid Jusélius Foundation. E.V.K. is supported by the Intramural Research Program of the US National Institutes of Health (NIH) National Library of Medicine. This work was supported in part through Laulima Government Solutions, LLC prime contract with the US 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. N.G.A.A. and S.F.E. are supported by the Agencia Estatal de Investigación (Spain) projects RTI2018-095700-B-I00 and PID2019-103998GB-I00, respectively. A.R.

M. is a Program Director at the US National Science Foundation (NSF), but the statements and opinions expressed herein are made in the personal capacity and do not constitute the endorsement by NSF or the government of the USA. E.M.A. is supported by the Biotechnology and Biological Sciences Research Council (BBSRC); under the BBSRC Institute Strategic Program Gut Microbes and Health BB/R012490/1 and its constituent projects BBS/E/F/000PR10353 and BBS/E/F/000PR10356. D.L.R. acknowledges support of the United Kingdom Medical Research Council (MC_UU_1201412). 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, US Department of Agriculture, Hatch Project, under Accession Number 1021494. J.M. is supported by Science Foundation Ireland under Grant Numbers 20/FFP-P/8664 and 15/SIRG/3430. N.V. acknowledges support in part by grants U01 AI151807 and R24 AI120942 from the US NIH. H.M.O. was supported by University of Helsinki funding for FINStruct and Instruct-ERIC research infrastructure. T.S. acknowledges support from the Austrian Science Fund (Grant Number P 28183). The work conducted by the US Department of Energy (DOE) Joint Genome Institute (S.R.), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. S.D. acknowledges support from the US National Science Foundation (DEB 1453241, OIA 1545553). B.E.D. was supported by the European Research Council (ERC) Consolidator grant 865694-DiversiPHI and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2051 – Project-ID 390713860, and the Alexander von Humboldt Foundation in the context of an Alexander von Humboldt-Professorship founded by German Federal Ministry of Education and Research. A.K. was supported by the ERC Consolidator grant 101001623-PALVIREVOL. J.M.K was supported by Deutsche Forschungsgemeinschaft within "Sonderforschungsbereich TRR 51". E.J.L. was supported by the NIH NIAID under Award Number U24AI162625. C.L. was supported by a Postdoctoral Mandate (PDMt2/21/038) from the KU Leuven Research Council and a fellowship (12D8623N) from the Research Foundation – Flanders (FWO). M.B.S. was supported by the US National Science Foundation awards OCE#1829831 and ABI#1759874. Y.B. was supported by the Professional Association of the Alliance of

United Kingdom, **38** Medical University of Vienna, Max Perutz Labs, Vienna Biocenter, Vienna, Austria, **39** Departments of Microbiology and Civil, Environmental, and Geodetic Engineering, Ohio State University, Columbus, Ohio, United States of America, **40** Institute of Plant Science and Resources, Okayama University, Kurashiki, Okayama, Japan, **41** School of Applied Sciences, College of Health, Science and Society, University of the West of England, Bristol, United Kingdom, **42** School of Animal and Comparative Biomedical Sciences, Department of Immunobiology, BIO5 Institute, and University of Arizona Cancer Center, Tucson, Arizona, United States of America, **43** KU Leuven, Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Leuven, Belgium, **44** Center for Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisbon, Portugal, **45** The Biodesign Center for Fundamental and Applied Microbiomics, School of Life Sciences, Center for Evolution and Medicine, Arizona State University, Tempe, Arizona, United States of America, **46** Department of Pathology, Center of Vector-Borne and Zoonotic Diseases, Institute for Human Infection and Immunity and World Reference Center for Emerging Viruses and Arboviruses, The University of Texas Medical Branch, Galveston, Texas, United States of America

* peter.simmonds@ndm.ox.ac.uk (PS); Evelien.Adriaenssens@quadram.ac.uk (EMA); zerbini@ufv.br (FMZ)

Abstract

A universal taxonomy of viruses is essential for a comprehensive view of the virus world and for communicating the complicated evolutionary relationships among viruses. However, there are major differences in the conceptualisation and approaches to virus classification and nomenclature among virologists, clinicians, agronomists, and other interested parties. Here, we provide recommendations to guide the construction of a coherent and comprehensive virus taxonomy, based on expert scientific consensus. Firstly, assignments of viruses should be congruent with the best attainable reconstruction of their evolutionary histories, i.e., taxa should be monophyletic. This fundamental principle for classification of viruses is currently included in the International Committee on Taxonomy of Viruses (ICTV) code only for the rank of species. Secondly, phenotypic and ecological properties of viruses may inform, but not override, evolutionary relatedness in the placement of ranks. Thirdly, alternative classifications that consider phenotypic attributes, such as being vector-borne (e.g., "arboviruses"), infecting a certain type of host (e.g., "mycoviruses," "bacteriophages") or displaying specific pathogenicity (e.g., "human immunodeficiency viruses"), may serve important clinical and regulatory purposes but often create polyphyletic categories that do not reflect evolutionary relationships. Nevertheless, such classifications ought to be maintained if they serve the needs of specific communities or play a practical clinical or regulatory role. However, they should not be considered or called taxonomies. Finally, while an evolution-based framework enables viruses discovered by metagenomics to be incorporated into the ICTV taxonomy, there are essential requirements for quality control of the sequence data used for these assignments. Combined, these four principles will enable future development and expansion of virus taxonomy as the true evolutionary diversity of viruses becomes apparent.

Introduction

The International Committee on Taxonomy of Viruses (ICTV) is the official body mandated by the International Union of Microbiology Societies to develop and maintain a taxonomy of viruses and the naming of their taxa. Throughout its history, the rules and codes associated with taxonomy have been updated many times in response to new discoveries, changes in

International Science Organizations (ANSO-PA-2020-07) and the Open Biodiversity and Health Big Data Programme of IUBS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: BSL, biosafety Level; ds, double-stranded; HCV, hepatitis C virus; HIV-1, human immunodeficiency virus type 1; HIV-2, HIV type 2; HTS, high-throughput sequencing; ICTV, International Committee on Taxonomy of Viruses; LIV, louping ill virus; LUCA, last universal cellular ancestor; MCP, major capsid protein; PDB, protein databank; RdRP, RNA-directed RNA polymerase; RT, reverse transcriptase; SARS-CoV, SARS coronavirus; ss, single-stranded; TBEV, tick-borne encephalitis virus.

understanding of evolutionary relationships among viruses, and, importantly, the advent of new technologies, such as high-throughput sequencing (HTS) that have vastly increased our global knowledge of viral diversity.

With roots in a pregenomic age, the criteria used for virus classification (see Box 1 for definitions of terms used in virus taxonomy) and taxon nomenclature were originally and by necessity based on observational properties of virus isolates, including the morphology of virion particles [1], type of nucleic acid in their genomes [2], and physical attributes such as susceptibility to inactivation by high temperature, organic solvents, and low pH [3,4]. While the vast majority of viruses now included in the ICTV taxonomy have been characterized at the genomic level (and this has been recently introduced as prerequisite for classification), there remains active debate on the extent to which historical reliance on physical and biological properties might continue to be useful as classification criteria and, indeed, whether viruses need to be characterized in in vitro culture or by virion visualization to be eligible for taxonomic assignment [5,6]. This topic is hotly debated among virologists, as among prokaryotic and fungal taxonomists, who are discussing whether to require strain isolation, phenotypic characterization, and placement in publicly available collections. Current prokaryote and fungi species lists capture only a small fraction of the true genetic diversity of these organisms in the wider environment, with species totals in the tens of thousands rather than the millions that genomic surveys estimate to exist [7,8].

Box 1. Definitions of terms used in virus taxonomy

Classification: The process of assigning viruses to groups. This process can be performed on different sets of features leading to different classification schemes. In the ICTV taxonomy, classification is evolutionarily based and hierarchical. The groups are named taxa.

Nomenclature: The naming of viruses or taxa. Taxon nomenclature is regulated by the ICTV and has a number of typographical restrictions concerning italicization and capitalization; taxon names above the rank of species possess suffixes to indicate taxonomic rank. Species nomenclature follows a binomial format (genus name + species epithet). In contrast, the naming of viruses is not regulated by the ICTV.

Rank: A relative position in a hierarchy. The ICTV taxonomy provides up to 15 ranks, with the highest (top) termed realm, and the lowest (basal) rank termed species.

Taxon: A taxonomic category for a group of viruses that is evolutionarily related and whose members may share similar properties. In a hierarchical classification, the demarcation criteria that define higher-rank taxa are shared with all lower-level taxa within.

Taxonomy: A biological classification based on evolutionary relationships in which viruses are assigned to a series of hierarchical taxa (classification) with regulated naming of component taxa (nomenclature).

An expert group convened by the ICTV in 2016 debated and affirmed a policy to allow viruses known from their genome sequences alone to be incorporated into virus taxonomy. This policy enables taxonomic assignments without requiring prior knowledge of a virus phenotypic properties, such as host range or pathogenicity, nor isolation of viruses in cell culture/

local lesion hosts, or visualization of virions [9]. Subsequent discussions led to the publication of guidelines for minimum standards for virus sequence data to ensure that viruses assigned to the ICTV taxonomy are represented by complete or coding-complete genomic sequences, which are accurately assembled and free from artifacts [10,11]. This development has led to large numbers of new taxa being incorporated into the official taxonomy, primarily from genomic data accrued from large-scale metagenomic surveys [12–18]. It also led to a renewed debate on the merits of having different criteria being used for taxonomic assignments among different groups of viruses. In particular, the emphasis on biological properties for many viruses infecting animals and plants versus the almost exclusive use of nucleic acid–based features for viruses infecting prokaryotes.

The creation of a unified evolutionary taxonomy that incorporates viruses classified both by traditional and metagenomics-based analyses requires considerable knowledge and insight into how virus properties are genomically encoded, about their evolutionary histories, and the influence of past recombination or reassortment of genomic regions on phylogenetic congruence. Furthermore, viruses have multiple, independent, and likely ancient evolutionary origins (reviewed in [15,19,20]). To develop criteria for assigning viruses to taxa, consensus is required on which genes are most informative in recovering relationships that best represent the evolutionary histories of each of these different clades.

Aims

A group of 45 basic and clinical virologists, bioinformaticians, and evolutionary and structural biologists met in Oxford, United Kingdom, in April 2022, to develop a community-wide consensus on methodologies used for virus classification and to establish an integrated and internally consistent taxonomic framework. The discussions focused primarily on how an evolutionary taxonomy of all viruses infecting eukaryotes, archaea, and bacteria might be constructed, which tools and approaches could be used, and how this process could be guided by identification of the most evolutionarily informative attributes of virus genome sequences and their organization. The group also considered the broader issue of how to reconcile an expanding genetic and structural classification with a partly phenetic classification developed by virologists over many decades that takes into account, among other properties, clinical and regulatory utility, virus/host ecology, and epidemiology.

The meeting achieved a substantial consensus on a range of approaches and challenges for taxonomy development, with all but two of the 45 participants endorsing a series of agreed recommendations in the form of four virus taxonomy principles (Box 2). We believe these will have long-term relevance and practical utility to inform the continued development of a universal virus taxonomy by the ICTV for many years to come.

Box 2. Recommendations for future virus taxonomy

1: Virus taxonomy should reflect the evolutionary history of viruses. Most viruses can be assigned to independent virus realms, each with an inferred separate evolutionary origin. Members of each realm possess sets of ancestral orthologous genes, termed hallmark genes, typically corresponding to replication or virion formation modules within their genomes. Their evolutionary relationships define monophyletic taxonomic assignments within each of these virus groups.

2: Virus properties may guide assignment of ranks to maximize their utility. While evolutionary relationships determine the topology of virus taxonomies, the ranks

assigned within it are human-made constructs, with up to 15 available from realm to species. Placement of viruses should follow patterns of evolutionary, genomic, and phenotypic properties; for example, species assignments may be based on host range, disease associations, or epidemiology, provided that such categories result in monophyletic groups.

3: Taxonomy is but one of many possible means to classify viruses. The taxonomy produced by the ICTV provides an overarching framework for classifying viruses based on evolutionary relationships. However, alternative classifications based on, for example, clinical or epidemiological properties or regulatory requirements have their own utilities in specific circumstances. These may not follow evolutionary relationships (like the Baltimore classification) or may include polyphyletic categories, such as arboviruses or human immunodeficiency viruses, that have epidemiological or clinical value but cannot be represented within an evolutionary taxonomy.

4: Taxonomic assignments of viruses inferred from metagenomic sequences require strict sequence quality control. Sequence-based assignment of a new taxon in the absence of other virus characterization requires it to be both accurate and complete. Published guidelines for minimum information about an uncultivated virus genome for taxonomic assignment have been produced [10].

Principle 1. Virus taxonomy should reflect the evolutionary history of viruses

Ranks used for virus taxonomy (realm, kingdom, phylum, class, order, family, genus, and species) must reflect degrees of evolutionary relatedness of the viruses assigned at each rank (Fig 1). This implies that viruses assigned to an individual rank form a monophyletic clade, i.e., all members of a rank share a most recent common ancestor that is distinct from all other evolutionary lineages assigned to the taxonomy despite the impact of gene acquisition, recombination, or reassortment events on genome organizations. This statement may seem obvious, but it is, in fact, the first formal recognition by the ICTV that virus taxonomy should be guided at all ranks by the inference of evolutionary history. This principle provides the necessary route forward for a taxonomy that can incorporate viruses characterized from metagenomics studies.

Virus evolutionary histories and choice of hallmark genes. We recognize that the establishment of a coherent virus taxonomy requires a variety of tools and approaches to reconstruct the underlying evolutionary relationships of viruses across their spectrum of diversity. Reconstruction of the deeper evolutionary histories of viruses is particularly challenging due to the lack of conserved genes across all virus genomes. This reflects the growing certainty that viruses have emerged on multiple independent occasions [20–22]. The impossibility of creating a taxonomic structure for all viruses with a single common ancestor contrasts with the biological classification of cellular life forms that possess a set of core genes, such as those encoding ribosomal proteins and ribosomal RNAs. While acknowledging the reticulate nature of the tree of life, these universal genes testify to the shared ancestry of genes present in bacteria, archaea, and eukaryotes linking back to a last universal cellular ancestor (LUCA) and that can be aligned to infer the deepest evolutionary relationships among all domains of cellular life forms [19,23].

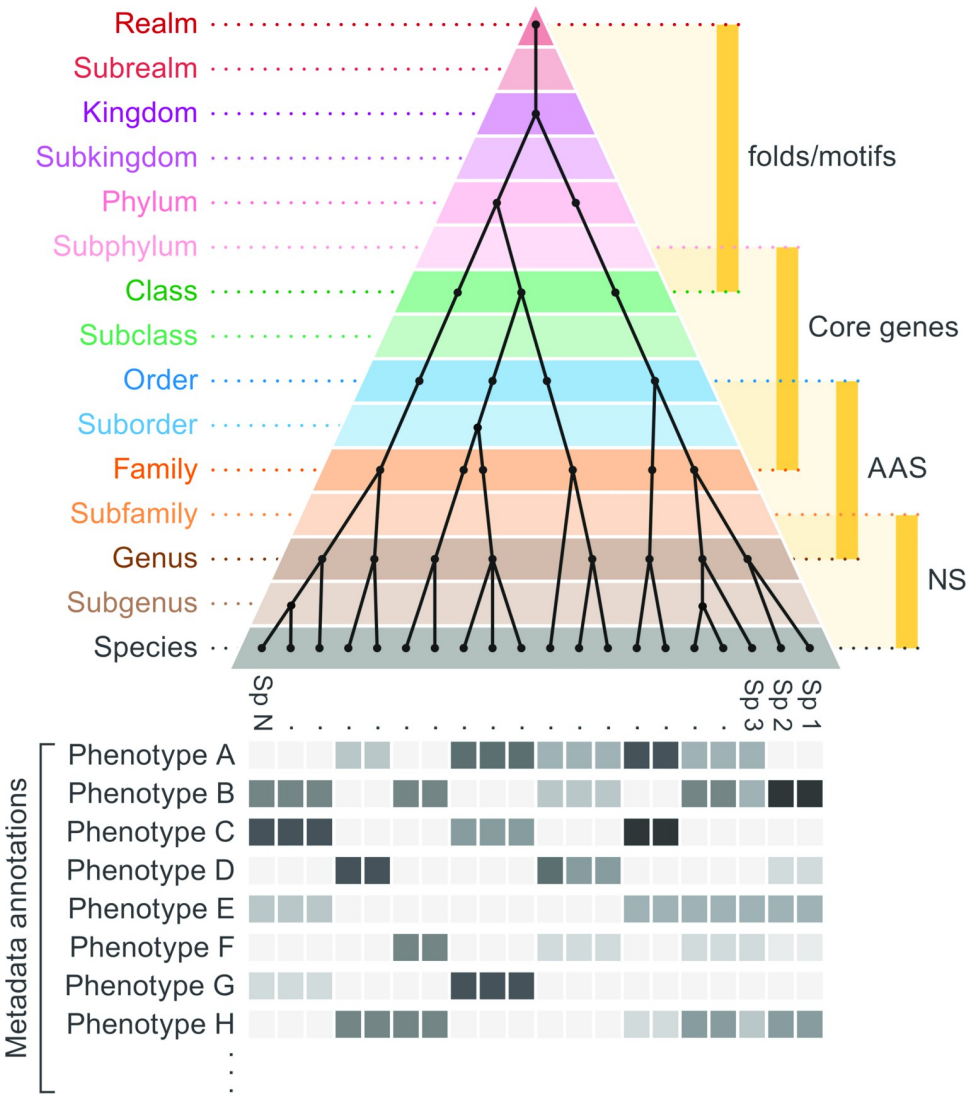


Fig 1. Ranks used in virus taxonomy. Schematic depiction of the 15-rank taxonomic framework used by the ICTV. It includes the methodologies that may be used to determine virus evolutionary relationships and make assignments at each rank. The pyramid shape indicates that the number of taxa increases from the top rank (realm) to the most basal rank (species, Sp.). The names of the 15 ranks are shown on the left of the pyramid, and the methodologies are on the right (AAS, amino acid sequence similarity; NS, nucleotide sequence similarity). The pyramid includes a hypothetical example of the taxonomy of a realm, indicating the number of taxa at each rank (filled circles). The phenotypic properties of classified viruses that may inform rank placements are depicted below the pyramid.

https://doi.org/10.1371/journal.pbio.3001922.g001

Despite the lack of universal virus genes, considerable progress has been made recently in better defining virus groups that share common ancestry [15,24]. The majority of viruses can be assigned to one of several independent realms, each of which is unified through possession of a shared orthologous gene or gene set, termed hallmark gene(s) [15]. Each realm is inferred to represent a distinct, independent origin of its constituent members. Two major functional components, the genome replication module and the virion formation module [25], are currently used for realm definition. These hallmark genes are thus considered to be ancestral to the members of each realm [25].

Virion morphogenesis modules were chosen as the defining characters for DNA viruses with larger genomes and govern assignments into the realms *Adnaviria*, *Duplodnaviria*, and

Varidnaviria. Viruses in these three realms encode major capsid proteins (MCPs) that are structurally radically different, as well as distinct virion assembly and genome packaging machineries [15,19,26], suggesting independent evolutionary origins. The evolutionary relationships of the genes involved in replication were not considered suitable for defining the realms of large DNA viruses because even relatively closely related viruses within the same realm often have distinct genome replication modules. For example, related viruses can encode nonhomologous or distantly related DNA polymerase genes of families A, B, or C that are interspersed with cellular counterparts. Some may lack DNA polymerase genes altogether and instead encode diverse replication initiators that facilitate the recruitment of the host replisome [25,27].

On the other hand, the key features of the genome replication machinery are the most suitable for defining the realms *Riboviria* [15], *Monodnaviria* (ICTV Taxonomy proposal 2019.005G.R.Monodnaviria), and *Ribozyviria* (ICTV Taxonomy proposal 2020.012D.R.Ribozyviria). The realm *Riboviria* unifies RNA viruses (kingdom *Orthornavirae*) and reverse-transcribing viruses (kingdom *Pararnavirae*), all of which encode homologous right-handed palm-domain RNA-directed RNA polymerase (RdRP) or reverse transcriptase (RT) genes, respectively. The phylogeny of these RdRPs and RTs was, therefore, used to guide the taxonomy within the *Riboviria*. In contrast, the capsid genes of RNA viruses fall into several unrelated groups, many likely to have been separately acquired from their hosts [28] or are completely absent. Analogously, all members of the realm *Monodnaviria* encode homologous histidine–hydrophobic residue–histidine (HUH) superfamily endonucleases [15,29], but the virion morphogenesis modules are distinct for viruses from different phyla within this realm. Finally, members of the *Kolmioviridae*, currently the sole family in the realm *Ribozyviria*, have small circular negative-sense RNA genomes that do not encode an RNA polymerase but contain a particular ribozyme that serves to define the realm.

Modular evolution of viruses. Virus evolution is frequently punctuated by large-scale genome reorganizations and the exchange of gene modules analogous to horizontal gene transfer in prokaryotes. For example, alpha-, beta-, gamma-, and deltaflexiviruses and tymoviruses possess an evolutionarily conserved set of replication genes (Rep) that define their classification in the order *Tymovirales* in the realm *Riboviria*. However, their capsid morphologies are diverse, including particles that are isometric (members of the *Tymoviridae*), filamentous/helical (viruses in the *Alphaflexiviridae*, *Betaflexiviridae*, and *Gammaflexiviridae*), or form no particles at all (members of the *Deltaflexiviridae* and fungus-infecting members of the *Alphaflexiviridae*). Even within a family, the phylogeny of capsid genes may be noncongruent with that of the replication genes, such as between genera of *Alphaflexiviridae* [30]. Similarly, members of the order *Martellivirales* share relatively closely related RdRPs and other genes involved in replication, such as helicases and capping enzymes, but produce flexible filamentous, rod-shaped, or icosahedral particles constructed from unrelated capsid proteins [28], or no classic virions at all (i.e., endornaviruses), suggesting the acquisition or loss of capsid morphogenesis genome modules from taxonomically distant viruses.

Furthermore, capsid genes can be exchanged between viruses that are otherwise evolutionarily unrelated. For example, a range of plant and animal RNA viruses and small single-stranded (ss) DNA viruses encode homologous horizontal single a jelly-roll capsid proteins, despite the RNA viruses being assigned to the realm *Riboviria* and the ssDNA viruses to the realm *Monodnaviria* [31–33]. Some prokaryotic viruses, in particular those alternating between lysogenic and lytic infections ("temperate" viruses), such as λ-like phages and those in the realm *Duplodnaviria* infecting *Mycolicibacterium* species, are substantially influenced by horizontal gene transfer [34]. These viruses possess a so-called mosaic genome structure, in which different parts of the genome can have quite different evolutionary histories [34]. In

such cases, the placement of taxonomic boundaries to form monophyletic groups at certain ranks is arbitrary as there are multiple possible evolutionary histories.

Although gene-sharing networks are informative for tracking gene exchange across virus groups [35], the relationships they depict violate the principles of ancestral descent that are used in taxonomy. Therefore, while different gene components are equally parts of the evolutionary histories of viruses and contribute to their phenotypes, for pragmatic purposes, we assign primacy to the most evolutionarily conserved hallmark genes in the construction of a hierarchical taxonomy. The use of hallmark genes for virus taxonomy is conceptually analogous to the use of a core set of conserved genes (primarily those for translation system components) for taxonomy of cellular life forms and eschews the use of the much more variable complements of genes subjected to horizontal gene transfer and loss [36]. Alternative taxonomies could be developed by selection of different genes to determine relatedness (for example, through basing the taxonomy of RNA viruses on capsid gene relationships, or of large DNA viruses by DNA polymerase genes). However, these typically yield a much greater number of unrelated virus groups and a less parsimonious association with virus properties.

Methodology for virus phylogenetics and taxonomy. Within individual virus realms, currently, a range of genome sequence comparison methods are needed to describe and assign viruses to different taxonomic ranks. For viruses with similar genome sequences, i.e., within the same species and genus, genetic relationships may be inferred from alignments of nucleotide or amino acid sequences of (near) complete genomes or of specific genes. The relationship among viruses can be further explored by phylogenetic tree inference and analysis, and where this is not practical, clustering by sequence similarity and analysis of pairwise distance distributions using tools such as PASC [37], DEmARC [38], and VIRIDIC [39]. However, these values only serve as an approximation of evolutionary relatedness [40,41]. The latter may be better inferred by phylogenetic methods that are also capable of calculating clade support, such as VICTOR [42] (Table 1).

Table 1. Examples of methodologies used for virus classification at different taxonomic ranks*.

Method	Principle	Rank range	Ref
DEmARC	Analysis of distributions of pairwise evolutionary distances between nucleotide or amino acid sequences	Suborder, Family, Subfamily, Genus, Subgenus, Species	[38]
GRAViTy	Virus relationships from composite Jaccard distances between HMM profiles and genome organizational models	Order, Family, (Genus)	[46]
HSF	Identification of structural equivalence and calculation of structural distances for structure-based phylogenetics	Realm, Kingdom, Phylum, Class, Order, Family, Genus	[58,68]
PASC	Analysis of pairwise nucleotide sequence distance distributions	Genus, Species	[37]
PhageClouds	Graph database of phage genomic sequences and intergenomic distances	Subfamily, Genus, Species	[69]
SDT	Pairwise nucleotide sequence alignment and identity calculation	Species	[70]
vConTACT2	Whole-genome gene sharing profiles integrating hierarchical clustering and confidence scores	Order, Family, Subfamily, Genus	[44,45]
VICTOR	Phylogenomic method optimized to ICTV classification that reports both sequence identity- and gene content-based phylogenies along with a suggested classification; works with either nucleotide or amino acid datasets	Family, Subfamily, Genus, Species	[42]
ViPTree	Virus relationships from genomic distances based translated nucleotide scores using tBLASTx	Family, Subfamily, Genus	[49]
VirClust	Hierarchical clustering based on core protein analysis	Order, Family, Subfamily	[71]
VIRIDIC	Calculates intergenomic similarities between pairs of viral genomes based on BLASTN alignments	Family, Genus, Species	[39]

*Discussions of tools dedicated to general reconstruction of phylogeny based on multiple sequence alignments are beyond the scope of this paper (for more information about this subject, see [72,73]). A more extensive list of virus bioinformatics tools including tools for virus taxonomy can be found at https://evirusbioinfc.notion.site/evirusbioinfc/18e21bc49827484b8a2f84463cb40b8d?v=92e7eb6703be4720abf17a901bc9a947.

https://doi.org/10.1371/journal.pbio.3001922.t001

At the intermediate levels of family, order, and class, relationships can be inferred by comparing sequences of evolutionarily conserved hallmark genes using sensitive methods for protein family profile comparison, such as HHPred [43], with subsequent phylogenetic analysis using appropriate methods for tree inference, for example, maximum likelihood methods. Comparison of hallmark proteins can be combined with metrics based on gene content, gene order and orientation (synteny), and other aspects of genome organization, using tools such as vConTACT2 [44,45] and GRAViTy [46], which are based on hierarchical clustering of gene sharing networks and the detection of hidden Markov model profiles of conserved protein families, with GRAViTy also taking into account metrics based on gene order and genome organization [47,48]. ViPTree [49] has been also used to define family level taxa of prokaryotic double-stranded (ds) DNA viruses, whereas VICTOR [42] can classify all prokaryotic viruses at the species, genus, subfamily, and family ranks through a joint clustering- and phylogeny-driven approach.

Taxonomic assignments at higher ranks, such as phyla, kingdoms, and realms, are based either on sequence comparison of the most highly conserved hallmark proteins and/or on protein structure comparisons. The latter can be informative for making evolutionary comparisons because homologous proteins typically retain similar structures, even when the corresponding amino acid sequences have diverged to the point that they are no longer sufficiently similar to infer homology based on sequence alone. Structure-based comparison methods include clustering based on estimates of distances between structures and structure-based phylogenetic analysis [50]. Much of the data used for this purpose originates from structures resolved experimentally with X-ray crystallography and, more recently, cryo-electron microscopy [51,52]. However, protein structure prediction methods have become much more accurate and insightful with the potential to enable large-scale bioinformatics-based reconstructions of structural features from sequence data alone. An important caveat is that, at this time, the recently developed and highly successful programs AlphaFold [53] and RosettaFold [54] generalize from known protein structures in the protein databank (PDB), a dataset in which virus proteins are substantially underrepresented [55], thus limiting their predictive power for analysis of relationships among viruses.

Hallmark gene-based assignments at the levels of kingdom and realm can be hampered by high levels of protein sequence divergence, with homology only detectable once high-resolution structures for the corresponding proteins become available. For this reason, the validity of the phylogenetic analyses used to designate kingdoms and phyla through evolutionary relationships among RdRP and RT genes of *Riboviria* using sequence analysis alone has been questioned, as these are based on the purported arbitrariness of the alignment of highly divergent sequences [56]. Alignment methodologies continue to be refined [57], but, ultimately, a range of sequence and protein structure comparison methods are likely to be required to delineate the higher ranks with confidence. Indeed, while protein structure can be influenced by environmental conditions (such as temperature, ionic strength, etc.), the optimal fold determined under standardized conditions is a highly evolutionarily conserved attribute of a protein coding sequence, and structural homology may be recoverable even when detectable sequence homology is lost. Encouragingly, a phylogeny based on protein structure comparisons of the viral RdRPs of members of *Riboviria* [58] matched the relationships inferred by aligned sequence comparison methods [15,59] at all but the highest ranks, as well as by the known functional diversification of these enzymes (i.e., transcription and priming mechanisms) and replication complex morphology [58,60].

Along similar lines, structure-based clustering and phylogeny of capsid proteins can provide a powerful approach when the reliable inference of evolutionary relationships by sequence comparisons ("traceability") is lost [61–63]. Thus, deeper evolutionary relationships that

underpin capsid protein structure and virion architecture may be used to classify large DNA viruses into realms and kingdoms. As an example, the structure-based PRD1-adenovirus lineage, whose members encode MCPs with a vertical double jelly-roll fold [61,63] (Fig 2), can be assigned to the kingdom *Bamfordvirae*, which falls within the realm *Varidnaviria*. Conversely, established structural relationships can now be used to inform sequence alignments and allow the incorporation of the ever-expanding wealth of virus sequence data into taxonomy [59,64].

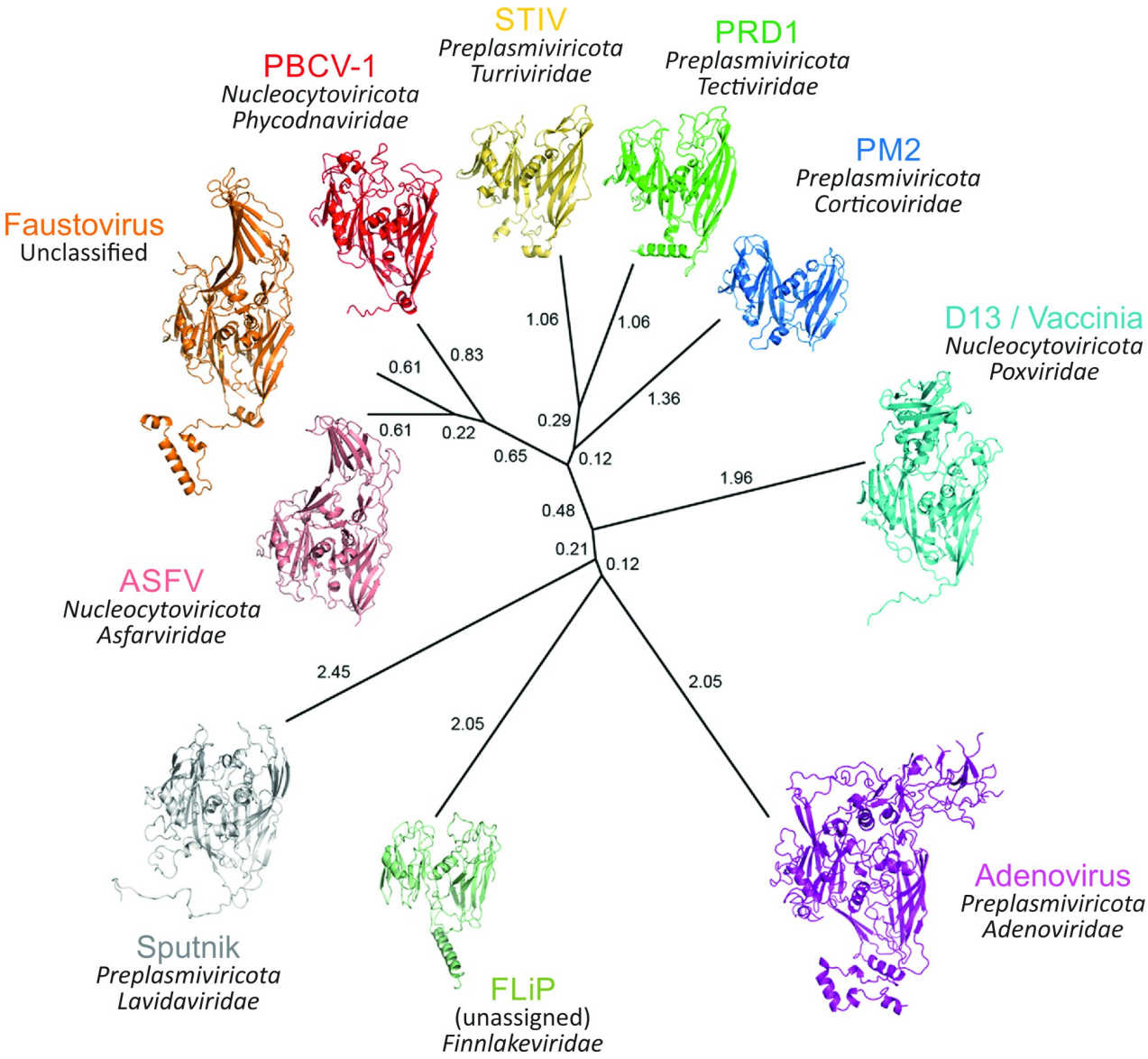


Fig 2. Structure-based dendrogram of capsid proteins of members of the kingdom *Bamfordvirae*. Structure-based phylogenetic tree inferred from major capsid protein (MCP) structures of the members of the kingdom *Bamfordvirae* in the *Varidnaviria* realm. Members of *Bamfordvirae* encode a vertical double-jelly roll fold MCP, which is the hallmark protein of this group of viruses. Next to each MCP structure are the virus name (top), the phylum (middle), and family (bottom), with "Faustovirus" not yet officially classified and *Finnlakeviridae* not yet assigned to any higher taxon. The evolutionary distances across the depicted members of the originally called PRD1-adenovirus viral lineage [67] were calculated with the Homologous Structure Finder software [50] and depicted with PHYLIP (https://evolution.genetics.washington.edu/phylip.html); the evolutionary distances are shown next to each branch. The protein data bank identifiers (PDBid) for the structures are as follows: PRD1: PDBid 1HX6; PBCV-1: 1M3Y; adenovirus: 1P2Z; STIV: 2BBD; Vaccinia D13: 2YGB; Sputnik: 3J26; Faustovirus: 5J7O; FLiP: 5OAC; ASFV p72: 6KU9; PM2: 2W0C. Adapted from [62].

https://doi.org/10.1371/journal.pbio.3001922.g002

Detection of subtle sequence conservation among structurally similar major capsid proteins of large DNA viruses further validates the use of these proteins as hallmarks for *Varidnaviria* and *Duplodnaviria* [64–66].

The ranges of sequence divergence (and, consequently, rank levels) over which the various analytical methods used in virus taxonomy are defined overlap substantially (Table 1). The recent delineation and assignment of a new family of bacterial viruses (*Herelleviridae*) [48] is an illustrative example of the value of such a combined approach. Concordance between multiple methods using different approaches increases the reliability of the taxonomic placement of novel taxa, whereas conflicts are informative regarding both the suitability of different comparison methods, and the nature of the relationships among viruses. Such conflicts can also arise from gene sharing networks and have led to several examples for which ICTV taxonomic revisions were needed [45]. Conflicts between different methods may also indicate the need to postpone taxonomic assignments until more data become available.

A six-realm taxonomy of viruses? Our understanding of virus origins and the evolutionary relationships inferred from hallmark gene trees within realms may change over time as analytical methods improve and new data become available. These may necessitate revisions to virus taxonomy. The extent to which the currently assigned realms encapsulate the full range of virus diversity and their distinct evolutionary origins remains under intense scrutiny. On the one hand, the possibility exists that large-scale metagenomics-based analyses of viruses in the environment are already approaching saturation of higher taxonomic ranks, such that the overall structure of virus taxonomy is stabilizing, even if many taxa remain to be delineated at lower levels. For example, a recent analysis of RNA virus diversity in the marine virome has vastly expanded the number of distinct viruses and putative genera and families, but these mostly can be assigned to the five previously established phyla within the riboviriad kingdom *Orthornavirae* [13]. On the other hand, the recent description of a plethora of RNA viruses sampled throughout the Global Oceans suggests the need to establish at least five additional phyla of RNA viruses [74] and urges some caution in these conclusions, particularly in light of the paucity of studies of virus diversity in other environments.

As an indication of future possible changes, structural analysis of virus capsid proteins within the realm *Varidnaviria* [75] indicates that this realm is not monophyletic and likely has to be split into two realms corresponding to the current kingdoms *Bamfordvirae* and *Helvetiavirae*. Furthermore, many (relatively) narrow groups of viruses, particularly those with hyperthermophilic archaea as hosts [76], cannot be classified into any of the six realms described to date. It appears highly likely that further characterization of the diversity of these virus groups and their conserved structural and genomic features will lead to the delineation of additional, comparatively small realms and the associated expansion of taxonomic ranks within these taxa [77].

While undoubtedly incomplete, the creation of the rank of realm and the recognition of a separate origin for each provides a substantive basis for a coherent and stable taxonomy of the viruses within them. The high value of hallmark gene relationships should be at the core of classification decisions at higher taxonomic ranks and will provide a blueprint for realm expansion, as needed in the future.

Principle 2. Virus properties may guide the assignment of ranks to maximize their utility

The primary value of taxonomy lies in its ability to sort organisms into categories that reflect their evolutionary history, be it at the species, genus, or family level, or higher ranks. All evolutionarily based taxonomic codes, including that of the ICTV, follow the principle that each

rank must be congruent with evolutionary relationships. The ICTV Code states that species should be monophyletic and, therefore, cannot group unrelated viruses sharing similarities in their physical properties, type of host/vector, disease associations, or other aspects of their phenotype, which might be polyphyletic in nature. This stipulation conflicts with many alternative (nontaxonomic) classifications of viruses used in medical, veterinary, agricultural, and regulatory fields described in Principle 3.

Assignment of taxonomic ranks. A hierarchical taxonomy based on evolutionary relationships of hallmark genes is inviolate and cannot support polyphyletic categories. However, the taxonomic rank is ultimately a human-made construct that arbitrarily assigns diversity into discrete categories that can be more readily conceptualized and named. The ICTV provides up to 15 ranks to partition virus diversity, from realm to species [78], although most viruses have historically been assigned to a more limited range, typically family, genus, and species. The placement of viruses at lower ranks of the taxonomy should follow patterns of natural clustering; virological knowledge and judgment are required to ensure, as far as possible, that placements also create informative categories for viruses with known phenotypic properties. Species assignments might therefore divide viruses based on their host range, disease associations, and epidemiological distributions, provided that such groups of viruses are monophyletic. The number of thresholds for delineating taxa at a given rank in the various virus groups would ideally be minimized to yield a more uniform taxonomy, although differences will remain, for example, between DNA and RNA viruses, which, in general, evolve at very different rates.

Such choices and the delineation of associated sequence divergence thresholds are typically made by expert groups of virologists, in most cases, ICTV Study Groups. As an example, the various genotypes of hepatitis C virus (HCV), all of which exclusively infect humans, were assigned by the ICTV *Flaviviridae* Study Group to the species *Hepacivirus C*, in distinction from those infecting New World monkey species (*Hepacivirus A* and *Hepacivirus B*) and horses (*Hepacivirus D*) [79]. This is possible because each group of host-associated viruses is monophyletic. The Study Group did not consider HCV genotypes, themselves each forming monophyletic groups, to represent separate species because they were not thought to be sufficiently different clinically or epidemiologically to merit such an assignment. This is despite their nucleotide sequence divergence (approximately 30%) being comparable to that between members of species assigned to other *Flaviviridae* genera, such as *Pestivirus A*, *Pestivirus B*, *Pestivirus C*, and *Pestivirus D*, which, in this case, show substantially different host ranges and disease associations. This constitutes one of many cases in which ecological drivers (in this case, host range) have shaped the evolutionary history of viruses and can be used to inform taxonomic outcomes.

Similarly, virological knowledge informed a revision of species demarcation criteria among members of the genus *Orthobunyavirus* (family *Peribunyaviridae*) [80]. The different geographical distributions, vector and host associations, and pathogenicities of Bunyamwera virus, Batai virus, Cache Valley virus, Ngari virus, Potosi virus, and Tensaw virus, all of which were originally assigned to a single species *Bunyamwera orthobunyavirus*, were considered to render these viruses so phenotypically distinctive as to warrant assignment to separate species [81]. The species nucleotide similarity thresholds demarcating species were accordingly increased to ensure that each of these distinctive viruses were assigned to different species, resulting in the reclassification of species within the genus (ICTV taxonomy proposal 2018.008M.A.v1.Orthobunyavirus_38sp).

The "usefulness" of a taxonomy combining clustering by sequence similarity with phenotypic properties highlights the need for extensive interdisciplinary work bridging the fields of bioinformatics and virology, creating and maintaining taxonomy as the knowledge of viral

diversity and their disease impact increases. At higher ranks, evolutionary and structural biologists may provide the more sophisticated approaches required to extract and evaluate genome sequence and structural features that depict the deepest evolutionary relationships of virus kingdoms and phyla.

Assessment of taxonomic ranks to viruses discovered in metagenomic studies. For viruses identified by genomes assembled in metagenomic sequence analyses, a phylogeny-based classification is required. Although information on the host range, effects on the host, and other phenotypic traits of these viruses is typically lacking, many of their attributes might be inferred from their genome organization and composition, evolutionary affinities of different genes, and base composition. In some cases, these characteristics may also suggest the likely host range [82–84]. Additional indications of the potential host range for these viruses can also be derived from their co-occurrence with specific groups of potential host organisms, as well as matches to CRISPR spacers in the case of viruses infecting bacteria and archaea [85–88]. As a follow-up to metagenomics, the properties of individual proteins or whole viruses can be experimentally determined through reverse genetics and characterization in vitro and in vivo, when possible (e.g., [89]). Nevertheless, in the (near) absence of phenotypic information, the taxonomy of these viruses, especially at the ranks below family, presents a challenge that might not be fully overcome until the genome analysis is complemented by virological studies at some point in the future.

There is, therefore, an asymmetry between the classification and rank assignments of viruses with well-defined phenotypic properties and the entirely genome-based classification required for viruses characterized by metagenomic analyses alone. Nevertheless, there is a clear consensus that viruses identified in metagenomic studies should be incorporated by the ICTV into virus taxonomy as far as the evidence allows [9]. Indeed, at this time and for the foreseeable future, this route provides by far the greatest source of information on genomic diversity of viruses, and recent metagenomics studies have led to the discovery of numerous new groups of viruses infecting hosts belonging to all three domains of cellular life. This has been a crucial advance in virology, and we must ensure that such viruses are incorporated into a taxonomy that will eventually catalog the true extent and complexity of the virosphere.

Principle 3. An evolutionary taxonomy is but one of many possible means to classify viruses

We recognize that classifications of viruses by clinicians, veterinarians, agronomists, and regulators may differ from a virus taxonomy that is grounded in evolutionary relationships. Numerous widely used clinical or veterinary virus designations cannot be supported by taxonomic assignments but often better serve clinical and regulatory purposes. The following cases, drawn from many possible examples, illustrate the various forms of mismatch that can occur.

Polyphyletic groups of viruses. The clinical and societal utility of the terms human immunodeficiency virus type 1 (HIV-1) and HIV type 2 (HIV-2) as the causative agents of AIDS requires no further explanation. However, the taxonomic assignment of these viruses has remained problematic because neither possesses a single common ancestor distinct from chimpanzee- (HIV-1) or sooty mangabey- (HIV-2) infecting viruses from which they derive (the phylogeny of HIV-1 is shown in Fig 3A) [90]. The only consistent ways to incorporate these viruses into a virus taxonomy based on evolutionary relationships are either to assign them as members of two species that each include lentiviruses infecting apes or Old World monkeys, or to designate each clade of HIV-1 (groups M, N, and O) and HIV-2 (groups A, B, P, and others) as separate species, these being distinct from multiple additional species of

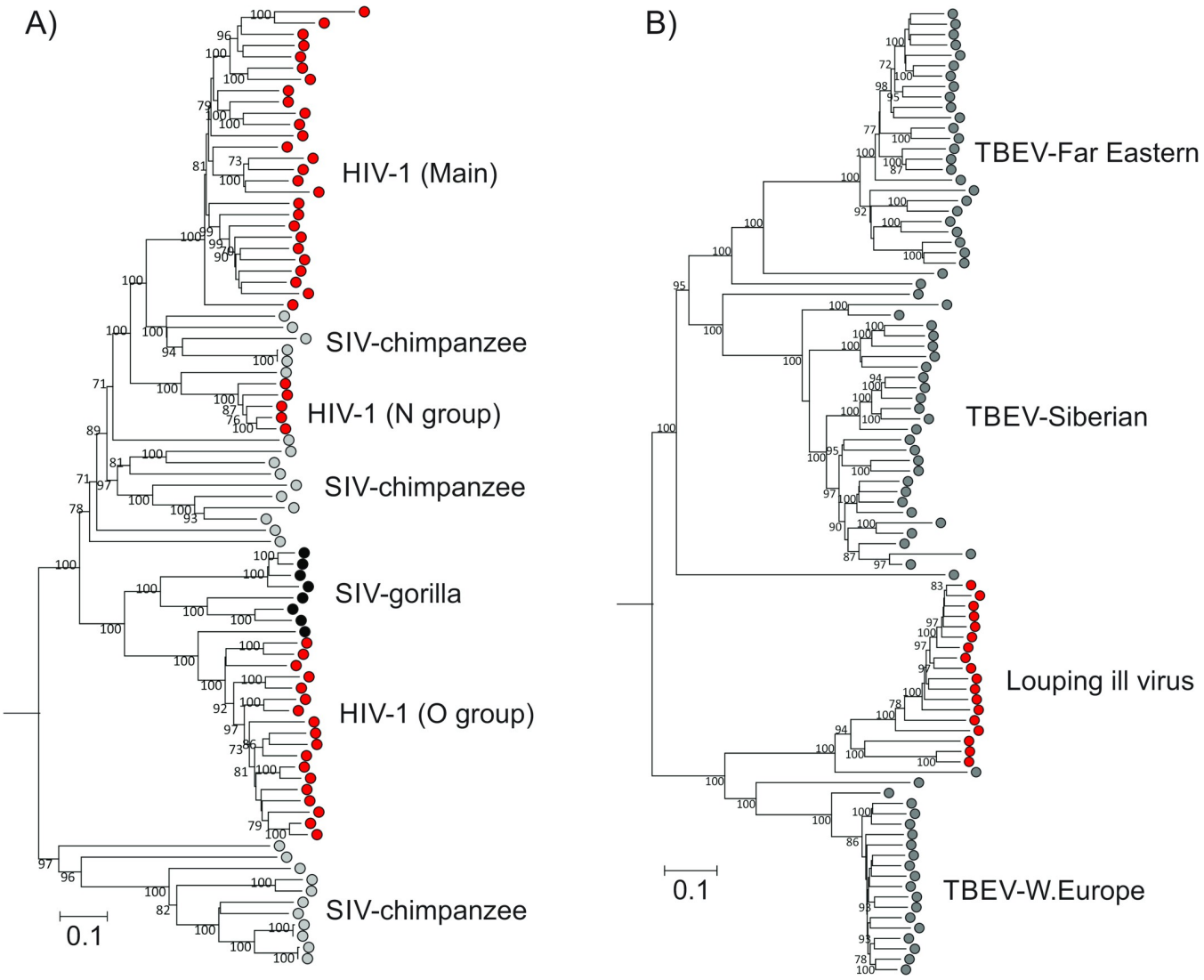


Fig 3. Examples of incompatibilities between species assignments and phylogenetic groupings. (**A**) Genetic relationships of HIV-1 (red dots) with simian immunodeficiency viruses infecting chimpanzees (gray dots) and gorillas (black dots). HIV-1 strains are polyphyletic and cannot be assigned to a single species taxon without incorporating nonhuman viruses within the definition. (**B**) Genetic relationships of louping ill virus (LIV) with tick-borne encephalitis virus (TBEV) strains isolated in Europe and Asia, with the principal groups labeled. Although LIV (red dots) is assigned to the species *Louping ill virus*, it lies within the phylogenetic tree created by strains of TBEV that all belong to the species *Tick-borne encephalitis virus*. The current assignment of LIV as a species therefore logically prevents strains of TBEV being assigned into a single species if species were to remain monophyletic. Trees were constructed from maximum composite likelihood distances between nucleotide sequences of (A) the *pol* gene of HIV-1/SIV and (B) the complete coding sequence of TBEV and LIV. To investigate the robustness of branches, nucleotide positions were bootstrap resampled 100 times as implemented in the MEGA7 program [91]; branches with 70% or greater support are labeled. The HIV-1/SIV tree was rooted using the HIV-2 sequence, M31113; the TBEV/LIV tree was rooted using the closely related Omsk haemorrhagic fever virus sequence, AY193805. Both trees have been annotated with a scale bar indicating substitutions per site.

https://doi.org/10.1371/journal.pbio.3001922.g003

simian viruses. However, neither of these potential taxonomies creates species that map directly onto the terms HIV-1 and HIV-2 or the collective term HIV. This discrepancy illustrates the need for separate and parallel classification of HIVs in clinical usage, distinct from their taxonomic classifications at the species level in the genus *Lentivirus*.

There is a parallel with the nomenclature and species assignments of SARS coronavirus (SARS-CoV) and SARS-CoV-2, the latter being the causative agent of COVID-19. As with HIV-1 and HIV-2, however, SARS-CoV and SARS-CoV-2 cannot logically be assigned to

separate virus species despite their clinical distinctiveness. They are classified collectively as members of the species *Severe acute respiratory syndrome-related virus*, along with a number of genetically closely sarbecoviruses infecting bats in South-East Asia [92]. While shared species membership follows from their genetic relatedness and inferred evolutionary origins, this taxonomic classification is incongruent with how the medical and wider community might want to classify them as agents of emerging and pandemic human infectious diseases.

Viruses that are assigned to different species but are not phylogenetically distinct. Louping ill virus (LIV) and tick-borne encephalitis virus (TBEV) are currently members of two separate species, *Louping ill virus* and *Tick-borne encephalitis virus*, respectively, in the genus *Flavivirus* of the family *Flaviviridae* [68]. The two viruses show distinct geographical distributions and host ranges; LIV is primarily present in sheep and grouse in the uplands of Scotland and South-West England and spread by the deer/sheep tick *Ixodes ricinus*, while TBEV is found in central Europe, Scandinavia, and large parts of Asia, with deer as the primary host reservoir. However, the current taxonomic assignments of LIV and TEBV to two separate species are not supported by their genetic relationships; members of the species *Louping ill virus* are phylogenetically interspersed with members of the species *Tick-borne encephalitis virus*. Consistent with previous observations [93], while sequences of LIVs are monophyletic, their common ancestor is not distinct from that of TBEV strains (Fig 3B), requiring either the assignment of both TBEV and LIV to a single species or the assignment of TBEV to several different species to reflect their distinct evolutionary histories. In neither case would these evolutionary taxonomic assignments reflect current widely used medical and veterinary terminology.

Members of the same species with distinct properties. The species *Enterovirus C*, in the family *Picornaviridae*, includes a clinically highly diverse range of member viruses, such as poliovirus types 1, 2, and 3, as well as several largely nonpathogenic enterovirus types. Their assignment to the same species was necessitated by their high degree of sequence similarity and their ability to recombine [94]. However, the poliovirus-associated neuroinvasive phenotype ultimately derives from a difference in the receptors used by these viruses, which is caused by only a handful of amino acid substitutions in the gene encoding the capsid protein VP1. Even though polioviruses are a fundamental element in disease descriptions (e.g., paralytic poliomyelitis and acute flaccid myelitis) and are targets of a largely successful global vaccination campaign, by evolutionary criteria, they cannot be classified into a species separate from many other enteroviruses, however appropriately that might reflect their clinical properties.

Virus groups described by phenetic attributes often do not map directly to taxa. Some broader terms such as "respiratory viruses," "viral meningitis," and "arboviruses" (arthropod-borne viruses) have wide clinical utility, being the staples of textbooks on infectious diseases and of medical reviews. However, each of the listed groups contains large sets of otherwise unrelated viruses across many different virus families and orders [95].

The Baltimore classes are incongruent with virus taxonomy. The classic paper by David Baltimore [96] proposed a classification of viruses based on the types of nucleic acids comprising the viral genome and the strategies used for genome replication and production of mRNA. The seven Baltimore classes, as they became known, have been widely adopted as an informal classification system for viruses. Although this classification system is logical and useful for understanding virus replication, it is at wide variance with evolutionary relationships of the viruses it classifies. Viruses in the realms *Adnaviria*, *Duplodnaviria*, most members of *Varidnaviria*, and some of *Monodnaviria* belong to Baltimore class I (dsDNA genomes), but other viruses in the latter two realms possess ssDNA genomes and accordingly belong to class II. Members of *Riboviria* with RNA genomes are represented in all the remaining classes III to VII, whereas the one member of the *Ribozyviria* is in class V [97].

Alternative classification of viruses. The broader point to be drawn from these examples is that an evolutionary taxonomy is not the only way to classify viruses, and its requirement to be congruent with evolutionary relationships can clash with classifications of viruses that are of greater value to clinicians, veterinarians, agronomists, regulators, and other stakeholders. It is similarly important to recognize that although species assignment thresholds can be selected so as to divide viruses into informative categories (see previous section), the requirement for congruency with evolutionary relationships means that this is not always possible.

The assignment of virus species with members often possessing quite distinct clinical or epidemiological attributes contrasts strongly with assignments of bacterial species in the classification of prokaryotes, for which each clinically or otherwise phenotypically distinct bacterial strain has been assigned to separate species with descriptive definitions. The situation is not unlike the historical classification practices of plant virologists; viruses were named after their specific disease presentations and assigned as unique members of a species bearing the same name. For example, a potyvirus causing mosaic in common bean was classified as a member of the species *Bean common mosaic virus*, whereas a second potyvirus causing mosaic in cowpea was classified as a member of *Blackeye cowpea mosaic virus*, and a third potyvirus causing dwarfing (growth reduction) in peanuts was classified as a member of *Peanut stripe virus*. Only when the genomes of these three viruses were sequenced did it become clear that they were closely related and therefore members of the same species, which currently retains the name *Bean common mosaic virus* [98].

Virus and species names. The ICTV has always maintained a typographic distinction between virus names and names of the taxa to which they are assigned [99], similar to how many other organisms have names distinct from the names of the species they are assigned to (e.g., humans ↔ *Homo sapiens*). Virus names are simply what virologists want to call viruses, inasmuch as naming practices have also evolved to address concerns such as discrimination and stigmatization (for example, by avoiding names based on geographical locations). Virus names are not within the remit of the ICTV. Viruses can be named with no restrictions on orthography (other than being non-italicized), numbering or language—indeed, many viruses have different names in different languages. Taxon names, in contrast, are within the mandate of the ICTV. They are written with an initial capital letter, are italicized, and may only contain letters of the Latin alphabet, Arabic numerals, and a limited number of symbols. Furthermore, taxon names are constant irrespective of the language that refers to them.

The ICTV typographic conventions reinforce the typological distinction between viruses as real-world objects and taxa as human-made classes or categories. This practice enables virus isolates to be mapped onto species assignments in a much more flexible way than the simple one-to-one correspondence that applies in many other areas of biology. This flexibility provides the framework for an evolutionarily based taxonomy to run alongside a variety of functional virus classifications without conflict. For example, the codes of practice for laboratory handling of viral pathogens are more useful if based upon a categorization of viruses, not species. This is exemplified by specific biocontainment requirements for HIV-1 and HIV-2, and the now highly restrictive regulatory framework established for poliovirus laboratory handling, which contrasts markedly with containment requirements for other members of the species *Enterovirus C*. Far Eastern and Russian spring–summer strains of TBEV are handled at Biosafety Level (BSL) 4 in the United States, whereas the European strains are at BSL 3 and LIV at BSL 2, regulatory distinctions that do not map onto their current species assignments.

Taxonomic definitions can incorporate elements of virus descriptions in their formulations, as is often the case with plant viruses. Similarly, virus descriptions can be more informative if they refer to the corresponding taxa to which they are assigned. Alternative classification systems have their own rules and utilities in the real world, and their use removes conflicts that

might otherwise arise if all virus classifications were irrevocably tied to an evolutionarily based taxonomy.

Principle 4. Taxonomic assignments of viruses inferred from metagenomic sequences require strict sequence quality control

As emphasized above, the assignment of viruses discovered by metagenomic analyses to new or established virus taxa must be based upon their genome sequences. Although the sample source is usually known, a genome sequence provides the key information about relatedness to other viruses, genome organization, inferred mechanisms of replication, and, increasingly, aspects of its virion structure, morphology, and even receptor use. As a unique source of information on the organism that is being classified, the genome sequence accordingly should be coding complete, i.e., contain the entire complement of protein-coding genes of the virus (as far as can be reasonably inferred), annotated and effectively free from sequencing or assembly errors. Thus, detailed bioinformatic information on the sequence acquisition methods used and their quality control is essential for taxonomic assignments of such viruses [10,11,100].

The ICTV does not require multiple, unique examples of sequences representing a new virus for taxonomic assignment, although characterization of additional members in such new taxa provides further information on genetic diversity and genome completeness. However, when a single sequence represents a species, it is particularly important to ensure that it depicts the virus genome as accurately and completely as possible [100]. The ICTV acknowledges the challenges for taxonomic classification of viruses discovered by metagenomics [11] and is currently working on specific guidelines for the submission of metagenomic sequences to public databases to facilitate taxonomic classification.

Acquisition of metagenomically derived sequence data in large environmental samples provides the best opportunity to fully explore and evaluate the true genetic diversity of viruses. However, the size and genetic complexity of such libraries and the widely used short read Illumina-based sequencing methods may hamper the assembly of complete genome sequences of viruses within samples. Consequently, much of the reported genetic diversity in such studies is based on phylogenetic comparisons of partial genome sequences, often restricted to reads spanning informative genes, such as the polymerases of riboviriad viruses. Such studies are vital for documenting the extent of virus diversity and, indeed, the completeness or otherwise of the current realm and kingdom structure of the virosphere. Partial sequences can also be used to improve the statistical support of large phylogenies that underlie the classification of viruses into taxonomic groups. However, our consensus view is that, without evidence of completeness, genome sequences obtained in such studies cannot be used as the sole basis for the creation of new taxa, meaning that the majority of viral sequences in current metagenomic datasets do not meet the standards for classification at this time. Further progress in long-read sequencing may help to speed up the acquisition of complete viral sequences with greater confidence in their proper assembly and completeness.

Conclusions

Our description of the four principles of virus taxonomy—which represent the consensus view of the workshop attendees—and the associated review of the evidence we provide gives a roadmap to the ICTV and its constituent expert committees and Study Groups to further develop and expand virus taxonomy. Implementation of the guidelines will also provide consistency and clarity for virus classification to the wider virology community. We acknowledge the vital contribution that expertise in bioinformatics and phylogenetics applied to virus sequence and structure analysis makes to the ever-expanding virus taxonomy. We also recognize the

importance of input from virology experts in developing a comprehensive view of the relationships among viruses at all taxonomic ranks.

While there is an established consensus on the need to incorporate viruses characterized in metagenomic studies into virus taxonomy [9], the outcomes of the 2022 workshop presented here effectively describe how this step can be achieved in practice. We have outlined the development and expansion of a taxonomy that was previously primarily based on disease and other phenotypically centered principles. We propose that virus taxonomy can and should now be based formally upon evolutionary relationships among viruses, with phenotypic properties being used where appropriate to inform the placement of lower ranks. This structure should enable the seamless incorporation of viruses characterized from their genomic sequences alone.

Acknowledgments

We thank Jiro Wada for excellent assistance with Fig 1 and Anya Crane for assistance with comprehensive editing of the manuscript.

The authors of this manuscript include all the members of the Executive Committee of the ICTV and a group of expert virologists who attended a 2022 workshop on virus taxonomy at Oxford University. The workshop was planned to discuss two main questions: "How can we unify the taxonomy of prokaryote viruses with the rest of viral taxonomy?" and "What are the informative characters used for virus taxonomy?" The discussion considered recent progress in prokaryote virus taxonomy, metagenomics, and bioinformatics. Although not exhaustive, the group of invited virologists was assembled to reflect a wide range of expertise in virus hosts, method development, viromics, structural biology, clinical and plant virology, as well as taking into account gender, career stage, and geographical diversity.

Author Contributions

Conceptualization: Peter Simmonds, Evelien M. Adriaenssens, F. Murilo Zerbini, Nicola G. A. Abrescia, Pakorn Aiewsakun, Poliane Alfenas-Zerbini, Yiming Bao, Jakub Barylski, Christian Drosten, Siobain Duffy, W. Paul Duprex, Bas E. Dutilh, Santiago F. Elena, Maria Laura García, Sandra Junglen, Aris Katzourakis, Eugene V. Koonin, Mart Krupovic, Jens H. Kuhn, Amy J. Lambert, Elliot J. Lefkowitz, Małgorzata Łobocka, Cédric Lood, Jennifer Mahony, Jan P. Meier-Kolthoff, Arcady R. Mushegian, Hanna M. Oksanen, Minna M. Poranen, Alejandro Reyes-Muñoz, David L. Robertson, Simon Roux, Luisa Rubino, Sead Sabanadzovic, Stuart Siddell, Tim Skern, Donald B. Smith, Matthew B. Sullivan, Nobuhiro Suzuki, Dann Turner, Koenraad Van Doorslaer, Anne-Mieke Vandamme, Arvind Varsani, Nikos Vasilakis.

Funding acquisition: Peter Simmonds.

Investigation: Peter Simmonds, Evelien M. Adriaenssens, F. Murilo Zerbini, Nicola G. A. Abrescia, Pakorn Aiewsakun, Poliane Alfenas-Zerbini, Yiming Bao, Jakub Barylski, Christian Drosten, Siobain Duffy, W. Paul Duprex, Bas E. Dutilh, Santiago F. Elena, Maria Laura García, Sandra Junglen, Aris Katzourakis, Eugene V. Koonin, Mart Krupovic, Jens H. Kuhn, Amy J. Lambert, Elliot J. Lefkowitz, Małgorzata Łobocka, Cédric Lood, Jennifer Mahony, Jan P. Meier-Kolthoff, Arcady R. Mushegian, Hanna M. Oksanen, Minna M. Poranen, Alejandro Reyes-Muñoz, David L. Robertson, Simon Roux, Luisa Rubino, Sead Sabanadzovic, Stuart Siddell, Tim Skern, Donald B. Smith, Matthew B. Sullivan, Nobuhiro Suzuki, Dann Turner, Koenraad Van Doorslaer, Anne-Mieke Vandamme, Arvind Varsani, Nikos Vasilakis.

Methodology: Evelien M. Adriaenssens, F. Murilo Zerbini.

Project administration: Peter Simmonds, Evelien M. Adriaenssens, F. Murilo Zerbini.

Resources: Evelien M. Adriaenssens, F. Murilo Zerbini.

Writing – original draft: Peter Simmonds.

Writing – review & editing: Peter Simmonds, Evelien M. Adriaenssens, F. Murilo Zerbini, Nicola G. A. Abrescia, Pakorn Aiewsakun, Poliane Alfenas-Zerbini, Yiming Bao, Jakub Barylski, Christian Drosten, Siobain Duffy, W. Paul Duprex, Bas E. Dutilh, Santiago F. Elena, Maria Laura García, Sandra Junglen, Aris Katzourakis, Eugene V. Koonin, Mart Krupovic, Jens H. Kuhn, Amy J. Lambert, Elliot J. Lefkowitz, Małgorzata Łobocka, Cédric Lood, Jennifer Mahony, Jan P. Meier-Kolthoff, Arcady R. Mushegian, Hanna M. Oksanen, Minna M. Poranen, Alejandro Reyes-Muñoz, David L. Robertson, Simon Roux, Luisa Rubino, Sead Sabanadzovic, Stuart Siddell, Tim Skern, Donald B. Smith, Matthew B. Sullivan, Nobuhiro Suzuki, Dann Turner, Koenraad Van Doorslaer, Anne-Mieke Vandamme, Arvind Varsani, Nikos Vasilakis.

References

1. Wildy P. Classifying viruses at higher levels: Symmetry and structure of virus particles as criteria. Symp Soc Gen Microbiol 1961; XII:145–63.
2. Cooper PD. A chemical basis for the classification of animal viruses. Nature. 1961; 190:302–305. https://doi.org/10.1038/190302a0 PMID: 13695331
3. Hamparian VV, Hilleman MR, Ketler A. Contributions to characterization and classification of animal viruses. Proc Soc Exp Biol Med. 1963; 112:1040–1050. https://doi.org/10.3181/00379727-112-28247 PMID: 13952431
4. Lwoff A, Tournier P. The classification of viruses. Annu Rev Microbiol. 1966; 20:45–74. https://doi.org/10.1146/annurev.mi.20.100166.000401 PMID: 5330240
5. Palmer M, Sutcliffe I, Venter SN, Hedlund BP. It is time for a new type of type to facilitate naming the microbial world. New Microbes New Infect. 2022; 47:100991. https://doi.org/10.1016/j.nmni.2022.100991 PMID: 35800027
6. Konstantinidis KT, Rosselló-Móra R, Amann R. Uncultivated microbes in need of their own taxonomy. ISME J. 2017; 11:2399–2406. https://doi.org/10.1038/ismej.2017.113 PMID: 28731467
7. Louca S, Mazel F, Doebeli M, Parfrey LW. A census-based estimate of Earth's bacterial and archaeal diversity. PLoS Biol. 2019; 17:e3000106. https://doi.org/10.1371/journal.pbio.3000106 PMID: 30716065
8. Gautam AK, Verma RK, Avasthi S, Sushma BY, Devadatha B, et al. Current insight into traditional and modern methods in fungal diversity estimates. J Fungi. 2022; 8:226. https://doi.org/10.3390/jof8030226 PMID: 35330228
9. Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, et al. Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol. 2017; 15:161–168. https://doi.org/10.1038/nrmicro.2016.177 PMID: 28134265
10. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). Nat Biotechnol. 2019; 37:29–37. https://doi.org/10.1038/nbt.4306 PMID: 30556814
11. Dutilh BE, Varsani A, Tong Y, Simmonds P, Sabanadzovic S, Rubino L, et al. Perspective on taxonomic classification of uncultivated viruses. Curr Opin Virol. 2021; 51:207–215. https://doi.org/10.1016/j.coviro.2021.10.011 PMID: 34781105
12. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, et al. Redefining the invertebrate RNA virosphere. Nature. 2016; 540:539–543. https://doi.org/10.1038/nature20167 PMID: 27880757
13. Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D, et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. Nat Microbiol. 2020; 5:1262–1270. https://doi.org/10.1038/s41564-020-0755-4 PMID: 32690954
14. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science. 2015; 348:1261498. https://doi.org/10.1126/science.1261498 PMID: 25999515

15. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, et al. Global organization and proposed megataxonomy of the virus world. Microbiol Mol Biol Rev. 2020; 84:e00061–e00019. https://doi.org/10.1128/MMBR.00061-19 PMID: 32132243

16. Roossinck MJ. Plant virus metagenomics: Biodiversity and ecology. Annu Rev Genet. 2012; 46:359–369. https://doi.org/10.1146/annurev-genet-110711-155600 PMID: 22934641

17. Li CX, Shi M, Tian JH, Lin XD, Kang YJ, Chen LJ, et al. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. elife. 2015; 4:e05378. https://doi.org/10.7554/eLife.05378 PMID: 25633976

18. Käfer S, Paraskevopoulou S, Zirkel F, Wieseke N, Donath A, Petersen M, et al. Re-assessing the diversity of negative strand RNA viruses in insects. PLoS Pathog. 2019; 15:e1008224. https://doi.org/10.1371/journal.ppat.1008224 PMID: 31830128

19. Krupovic M, Dolja VV, Koonin EV. The LUCA and its complex virome. Nat Rev Microbiol. 2020; 18:661–670. https://doi.org/10.1038/s41579-020-0408-x PMID: 32665595

20. Nasir A, Romero-Severson E, Claverie JM. Investigating the concept and origin of viruses. Trends Microbiol. 2020; 28:959–967. https://doi.org/10.1016/j.tim.2020.08.003 PMID: 33158732

21. Brussow H. The not so universal Tree of Life or the place of viruses in the living world. Philos Trans R Soc Lond Ser B Biol Sci. 2009; 364:2263–2274. https://doi.org/10.1098/rstb.2009.0036 PMID: 19571246

22. Krupovic M, Dolja VV, Koonin EV. Origin of viruses: primordial replicators recruiting capsids from hosts. Nat Rev Microbiol. 2019; 17:449–458. https://doi.org/10.1038/s41579-019-0205-6 PMID: 31142823

23. Woese C. The universal ancestor. Proc Natl Acad Sci U S A. 1998; 95:6854–6859. https://doi.org/10.1073/pnas.95.12.6854 PMID: 9618502

24. Kuhn JH, Wolf YI, Krupovic M, Zhang YZ, Maes P, Dolja VV, et al. Classify viruses—the gain is worth the pain. Nature. 2019; 566:318–320. https://doi.org/10.1038/d41586-019-00599-8 PMID: 30787460

25. Krupovic M, Bamford DH. Order to the viral universe. J Virol. 2010; 84:12476–12479. https://doi.org/10.1128/JVI.01489-10 PMID: 20926569

26. Krupovic M, Kuhn JH, Wang F, Baquero DP, Dolja VV, Egelman EH, et al. *Adnaviria:* a new realm for archaeal filamentous viruses with linear A-form double-stranded DNA genomes. J Virol. 2021; 95: e0067321. https://doi.org/10.1128/JVI.00673-21 PMID: 34011550

27. Kazlauskas D, Krupovic M, Venclovas Č. The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. Nucleic Acids Res. 2016; 44:4551–4564. https://doi.org/10.1093/nar/gkw322 PMID: 27112572

28. Krupovic M, Koonin EV. Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A. 2017; 114:E2401–e10. https://doi.org/10.1073/pnas.1621061114 PMID: 28265094

29. Kazlauskas D, Varsani A, Koonin EV, Krupovic M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. Nat Commun. 2019; 10:3425.

30. Kreuze JF, Vaira AM, Menzel W, Candresse T, Zavriev SK, Hammond J, et al. ICTV Virus Taxonomy Profile: *Alphaflexiviridae*. J Gen Virol. 2020; 101:699–700. https://doi.org/10.1099/jgv.0.001436 PMID: 32525472

31. Roux S, Enault F, Bronner G, Vaulot D, Forterre P, Krupovic M. Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. Nat Commun. 2013; 4:2700. https://doi.org/10.1038/ncomms3700 PMID: 24193254

32. de la Higuera I, Kasun GW, Torrance EL, Pratt AA, Maluenda A, Colombet J, et al. Unveiling crucivirus diversity by mining metagenomic data. MBio. 2020; 11:e01410–e01420. https://doi.org/10.1128/mBio.01410-20 PMID: 32873755

33. Kazlauskas D, Dayaram A, Kraberger S, Goldstien S, Varsani A, Krupovic M. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. Virology. 2017; 504:114–121. https://doi.org/10.1016/j.virol.2017.02.001 PMID: 28189969

34. Mavrich TN, Hatfull GF. Bacteriophage evolution differs by host, lifestyle and genome. Nat Microbiol. 2017; 2:17112. https://doi.org/10.1038/nmicrobiol.2017.112 PMID: 28692019

35. Shapiro JW, Putonti C. Gene co-occurrence networks reflect bacteriophage ecology and evolution. MBio. 2018; 9:e01870–e01817. https://doi.org/10.1128/mBio.01870-17 PMID: 29559574

36. Doolittle WF. Phylogenetic classification and the universal tree. Science. 1999; 284:2124–2129. https://doi.org/10.1126/science.284.5423.2124 PMID: 10381871

37. Bao Y, Chetvernin V, Tatusova T. Improvements to pairwise sequence comparison (PASC): A genome-based web tool for virus classification. Arch Virol. 2014; 159:3293–3304. https://doi.org/10.1007/s00705-014-2197-x PMID: 25119676

38. Lauber C, Gorbalenya AE. Toward genetics-based virus taxonomy: comparative analysis of a genetics-based classification and the taxonomy of picornaviruses. J Virol. 2012; 86:3905–3915. https://doi.org/10.1128/JVI.07174-11 PMID: 22278238

39. Moraru C, Varsani A, Kropinski AM. VIRIDIC—A novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. Viruses. 2020; 12:1268. https://doi.org/10.3390/v12111268 PMID: 33172115

40. Eisen JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 1998; 8:163–167. https://doi.org/10.1101/gr.8.3.163 PMID: 9521918

41. Smith SA, Pease JB. Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny. Brief Bioinform. 2017; 18:451–457. https://doi.org/10.1093/bib/bbw034 PMID: 27103098

42. Meier-Kolthoff JP, Göker M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics. 2017; 33:3396–3404. https://doi.org/10.1093/bioinformatics/btx440 PMID: 29036289

43. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005; 33:W244–W248. https://doi.org/10.1093/nar/gki408 PMID: 15980461

44. Bolduc B, Jang HB, Doulcier G, You ZQ, Roux S, Sullivan MB. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea and Bacteria*. PeerJ. 2017; 5:e3243. https://doi.org/10.7717/peerj.3243 PMID: 28480138

45. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. 2019; 37:632–639. https://doi.org/10.1038/s41587-019-0100-8 PMID: 31061483

46. Aiewsakun P, Simmonds P. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. Microbiome. 2018; 6:38. https://doi.org/10.1186/s40168-018-0422-7 PMID: 29458427

47. Liu Y, Demina TA, Roux S, Aiewsakun P, Kazlauskas D, Simmonds P, et al. Diversity, taxonomy, and evolution of archaeal viruses of the class *Caudoviricetes*. PLoS Biol. 2021; 19:e3001442. https://doi.org/10.1371/journal.pbio.3001442 PMID: 34752450

48. Barylski J, Enault F, Dutilh BE, Schuller MBP, Edwards RA, Gillis A, et al. Analysis of spounaviruses as a case study for the overdue reclassification of tailed phages. Syst Biol. 2019:110–23.

49. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. ViPTree: the viral proteomic tree server. Bioinformatics. 2017; 33:2379–2380. https://doi.org/10.1093/bioinformatics/btx157 PMID: 28379287

50. Ravantti J, Bamford D, Stuart DI. Automatic comparison and classification of protein structures. J Struct Biol. 2013; 183:47–56. https://doi.org/10.1016/j.jsb.2013.05.007 PMID: 23707633

51. Binshtein E, Ohi MD. Cryo-electron microscopy and the amazing race to atomic resolution. Biochemistry. 2015; 54:3133–3141. https://doi.org/10.1021/acs.biochem.5b00114 PMID: 25955078

52. Santos-Pérez I, Charro D, Gil-Carton D, Azkargorta M, Elortza F, Bamford DH, et al. Structural basis for assembly of vertical single β-barrel viruses. Nat Commun. 2019; 10:1184.

53. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020; 577:706–710. https://doi.org/10.1038/s41586-019-1923-7 PMID: 31942072

54. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021; 373:871–876. https://doi.org/10.1126/science.abj8754 PMID: 34282049

55. Robson B. Testing machine learning techniques for general application by using protein secondary structure prediction. A brief survey with studies of pitfalls and benefits using a simple progressive learning approach. Comput Biol Med. 2021; 138:104883. https://doi.org/10.1016/j.compbiomed.2021.104883 PMID: 34598067

56. Holmes EC, Duchêne S. Can sequence phylogenies safely infer the origin of the global virome? mBio. 2019; 10:e00289–e00219. https://doi.org/10.1128/mBio.00289-19 PMID: 30992348

57. Edgar RC. MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping. bioRxiv. 2021:2021.06.20.449169.

58. Mönttinen HAM, Ravantti JJ, Poranen MM. Structure unveils relationships between RNA virus polymerases. Viruses. 2021; 13:313. https://doi.org/10.3390/v13020313 PMID: 33671332

59. Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, et al. Origins and evolution of the global RNA virome. MBio. 2018; 9:e02329–e02318. https://doi.org/10.1128/mBio.02329-18 PMID: 30482837

60. Ahola T. New phylogenetic grouping of positive-sense RNA viruses is concordant with replication complex morphology. MBio. 2019; 10:e01402–e01419. https://doi.org/10.1128/mBio.01402-19 PMID: 31363030

61. Abrescia NG, Bamford DH, Grimes JM, Stuart DI. Structure unifies the viral universe. Annu Rev Biochem. 2012; 81:795–822. https://doi.org/10.1146/annurev-biochem-060910-095130 PMID: 22482909

62. Ravantti JJ, Martinez-Castillo A, Abrescia NGA. Superimposition of viral protein structures: A means to decipher the phylogenies of viruses. Viruses. 2020; 12:1146. https://doi.org/10.3390/v12101146 PMID: 33050291

63. Krupovic M, Bamford DH. Virus evolution: how far does the double beta-barrel viral lineage extend? Nat Rev Microbiol. 2008; 6:941–948. https://doi.org/10.1038/nrmicro2033 PMID: 19008892

64. Yutin N, Bäckström D, Ettema TJG, Krupovic M, Koonin EV. Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis. Virol J. 2018; 15:67. https://doi.org/10.1186/s12985-018-0974-y PMID: 29636073

65. Sinclair RM, Ravantti JJ, Bamford DH. Nucleic and amino acid sequences support structure-based viral classification. J Virol. 2017; 91:e02275–e02216. https://doi.org/10.1128/JVI.02275-16 PMID: 28122979

66. Van Doorslaer K, Ruoppolo V, Schmidt A, Lescroel A, Jongsomjit D, Elrod M, et al. Unique genome organization of non-mammalian papillomaviruses provides insights into the evolution of viral early proteins. Virus Evol 2017; 3:vex027. https://doi.org/10.1093/ve/vex027 PMID: 29026649

67. Bamford DH, Grimes JM, Stuart DI. What does structure tell us about virus evolution? Curr Opin Struct Biol. 2005; 15:655–663. https://doi.org/10.1016/j.sbi.2005.10.012 PMID: 16271469

68. Gao GF, Zanotto PM, Holmes EC, Reid HW, Gould EA. Molecular variation, evolution and geographical distribution of louping ill virus. Acta Virol. 1997; 41:259–268. PMID: 9607079

69. Rangel-Pineros G, Millard A, Michniewski S, Scanlan D, Sirén K, Reyes A, et al. From trees to clouds: PhageClouds for fast comparison of ~640,000 phage genomic sequences and host-centric visualization using genomic network graphs. Phage. 2021; 2:194–203.

70. Muhire BM, Varsani A, Martin DP. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS ONE. 2014; 9:e108277. https://doi.org/10.1371/journal.pone.0108277 PMID: 25259891

71. Moraru C. VirClust–a tool for hierarchical clustering, core gene detection and annotation of (prokaryotic) viruses. bioRxiv. 2021:2021.06.14.448304.

72. Low SJ, Džunková M, Chaumeil PA, Parks DH, Hugenholtz P. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. Nat Microbiol. 2019; 4:1306–1315. https://doi.org/10.1038/s41564-019-0448-z PMID: 31110365

73. Gorbalenya AE, Lauber C. Bioinformatics of virus taxonomy: foundations and tools for developing sequence-based hierarchical classification. Curr Opin Virol. 2022; 52:48–56. https://doi.org/10.1016/j.coviro.2021.11.003 PMID: 34883443

74. Zayed AA, Wainaina JM, Dominguez-Huerta G, Pelletier E, Guo J, Mohssen M, et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. Science. 2022; 376:156–162. https://doi.org/10.1126/science.abm5847 PMID: 35389782

75. Krupovic M, Makarova KS, Koonin EV. Cellular homologs of the double jelly-roll major capsid proteins clarify the origins of an ancient virus kingdom. Proc Natl Acad Sci U S A. 2022; 119:e2120620119. https://doi.org/10.1073/pnas.2120620119 PMID: 35078938

76. Baquero DP, Liu Y, Wang F, Egelman EH, Prangishvili D, Krupovic M. Structure and assembly of archaeal viruses. Adv Virus Res. 2020; 108:127–164. https://doi.org/10.1016/bs.aivir.2020.09.004 PMID: 33837715

77. Koonin EV, Krupovic M, Dolja VV. The global virome: How much diversity and how many independent origins? Environ Microbiol. 2022. https://doi.org/10.1111/1462-2920.16207 PMID: 36097140

78. Gorbalenya AE. Increasing the number of available ranks in virus taxonomy from five to ten and adopting the Baltimore classes as taxa at the basal rank. Arch Virol. 2018; 163:2933–2936. https://doi.org/10.1007/s00705-018-3915-6 PMID: 29942981

79. Simmonds P, Becher P, Bukh J, Gould EA, Meyers G, Monath T, et al. ICTV Virus Taxonomy Profile: *Flaviviridae*. J Gen Virol. 2017; 98:2–3. https://doi.org/10.1099/jgv.0.000672 PMID: 28218572

80. Hughes HR, Adkins S, Alkhovskiy S, Beer M, Blair C, Calisher CH, et al. ICTV Virus Taxonomy Profile: *Peribunyaviridae*. J Gen Virol. 2020; 101:1–2. https://doi.org/10.1099/jgv.0.001365 PMID: 31846417

81. Blitvich BJ, Beaty BJ, Blair CD, Brault AC, Dobler G, Drebot MA, et al. Bunyavirus taxonomy: Limitations and misconceptions associated with the current ICTV criteria used for species demarcation. Am J Trop Med Hyg. 2018; 99:11–16. https://doi.org/10.4269/ajtmh.18-0038 PMID: 29692303

82. Lee B, Smith DK, Guan Y. Alignment free sequence comparison methods and reservoir host prediction. Bioinformatics. 2021; 37:3337–3342. https://doi.org/10.1093/bioinformatics/btab338 PMID: 33964132

83. Babayan SA, Orton RJ, Streicker DG. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. Science. 2018; 362:577–580. https://doi.org/10.1126/science.aap9072 PMID: 30385576

84. Zielezinski A, Girgis HZ, Bernard G, Leimeister CA, Tang K, Dencker T, et al. Benchmarking of alignment-free sequence comparison methods. Genome Biol. 2019; 20:144. https://doi.org/10.1186/s13059-019-1755-7 PMID: 31345254

85. Lu C, Zhang Z, Cai Z, Zhu Z, Qiu Y, Wu A, et al. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. BMC Biol. 2021; 19:5. https://doi.org/10.1186/s12915-020-00938-6 PMID: 33441133

86. Sugimoto R, Nishimura L, Nguyen PT, Ito J, Parrish NF, Mori H, et al. Comprehensive discovery of CRISPR-targeted terminally redundant sequences in the human gut metagenome: Viruses, plasmids, and more. PLoS Comput Biol. 2021; 17:e1009428. https://doi.org/10.1371/journal.pcbi.1009428 PMID: 34673779

87. Coclet C, Roux S. Global overview and major challenges of host prediction methods for uncultivated phages. Curr Opin Virol. 2021; 49:117–126. https://doi.org/10.1016/j.coviro.2021.05.003 PMID: 34126465

88. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiol Rev. 2016; 40:258–272. https://doi.org/10.1093/femsre/fuv048 PMID: 26657537

89. Lauber C, Seitz S, Mattei S, Suh A, Beck J, Herstein J, et al. Deciphering the origin and evolution of hepatitis B viruses by means of a family of non-enveloped fish viruses. Cell Host Microbe. 2017; 22:387–99.e6. https://doi.org/10.1016/j.chom.2017.07.019 PMID: 28867387

90. Coffin J, Blomberg J, Fan H, Gifford R, Hatziioannou T, Lindemann D, et al. ICTV Virus Taxonomy Profile: *Retroviridae* 2021. J Gen Virol. 2021; 102(12):001712. https://doi.org/10.1099/jgv.0.001712 PMID: 34939563

91. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016; 33:1870–1874. https://doi.org/10.1093/molbev/msw054 PMID: 27004904

92. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. The species *Severe acute respiratory syndrome-related coronavirus:* classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020; 5:536–544. https://doi.org/10.1038/s41564-020-0695-z PMID: 32123347

93. Bondaryuk AN, Andaev EI, Dzhioev YP, Zlobin VI, Tkachev SE, Kozlova IV, et al. Delimitation of the tick-borne flaviviruses. Resolving the tick-borne encephalitis virus and louping-ill virus paraphyletic taxa. Mol Phylogenet Evol. 2022; 169:107411. https://doi.org/10.1016/j.ympev.2022.107411 PMID: 35032647

94. Brown B, Oberste MS, Maher K, Pallansch MA. Complete genomic sequencing shows that polioviruses and members of human enterovirus species C are closely related in the noncapsid coding region. J Virol. 2003; 77:8973–8984. https://doi.org/10.1128/jvi.77.16.8973-8984.2003 PMID: 12885914

95. Kuhn JH, Crozier I. Arthropod-borne and rodent-borne virus infections. In: Loscalzo J, Fauci AS, Kasper DL, Hauser SL, Longo DL, Jameson JL, editors. Harrison's Principles of Internal Medicine. 2. 21 ed. Columbus, Ohio, USA: McGraw-Hill Education; 2022. p. 1624–545.

96. Baltimore D. Expression of animal virus genomes. Bacteriol Rev. 1971; 35:235–241. https://doi.org/10.1128/br.35.3.235-241.1971 PMID: 4329869

97. Koonin EV, Krupovic M, Agol VI. The Baltimore classification of viruses 50 years later: How Does it stand in the light of virus evolution? Microbiol Mol Biol Rev. 2021; 85:e0005321. https://doi.org/10.1128/MMBR.00053-21 PMID: 34259570

98. Wylie SJ, Adams M, Chalam C, Kreuze J, López-Moya JJ, Ohshima K, et al. ICTV Virus Taxonomy Profile: *Potyviridae*. J Gen Virol. 2017; 98:352–354. https://doi.org/10.1099/jgv.0.000740 PMID: 28366187

99. Zerbini FM, Siddell SG, Mushegian AR, Walker PJ, Lefkowitz EJ, Adriaenssens EM, et al. Differentiating between viruses and virus species by writing their names correctly. Arch Virol. 2022; 167:1231–1234. https://doi.org/10.1007/s00705-021-05323-4 PMID: 35043230

100. Ladner JT, Beitzel B, Chain PS, Davenport MG, Donaldson EF, Frieman M, et al. Standards for sequencing viral genomes in the era of high-throughput sequencing. MBio. 2014; 5:e01360–e01314. https://doi.org/10.1128/mBio.01360-14 PMID: 24939889