Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Virology

Perspective on taxonomic classification of uncultivated viruses

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

Historically, virus taxonomy has been limited to describing viruses that were readily cultivated in the laboratory or emerging in natural biomes. Metagenomic analyses, single-particle sequencing, and database mining efforts have yielded new sequence data on an astounding number of previously unknown viruses. As metagenomes are relatively free of biases, these data provide an unprecedented insight into the vastness of the virosphere, but to properly value the extent of this diversity it is critical that the viruses are taxonomically classified. Inclusion of uncultivated viruses has already improved the process as well as the understanding of the taxa, viruses, and their evolutionary relationships. The continuous development and testing of computational tools will be required to maintain a dynamic virus taxonomy that can accommodate the new discoveries.

Addresses
[1] Theoretical Biology and Bioinformatics, Science for Life, Utrecht University, Padualaan 8, 3584 CH, Utrecht, The Netherlands
[2] Institute of Bioloversity, Faculty of Biological Sciences, Cluster of Excellence Balance of the Microverse, Friedrich-Schiller-University Jena, 07743, Jena, Germany
[3] The Biodesign Center of Fundamental and Applied Microbiomics, School of Life Sciences, Center for Evolution and Medicine, Arizona State University, Tempe, AZ 85287, USA
[4] Structural Biology Research Unit, Department of Integrative Biomedical Sciences, University of Cape Town, 7925, Cape Town, South Africa
[5] Beijing Advanced Innovation Centre for Soft Matter Science and Engineering, College of Life Science and Technology, Beijing University of Chemical Technology, Beijing, 100029, China
[6] Nuffield Department of Medicine, University of Oxford, Peter Medawar Building, South Parks Road, Oxford, OX1 3SY, UK
[7] Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, MS 39762, USA
[8] Istituto per la Protezione Sostenibile delle Piante, Consiglio Nazionale delle Ricerche, Bari, Italy
[9] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[10] Max Planck Tandem Group in Computational Biology, Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia
[11] Department of Microbial and Molecular Systems, KU Leuven, Kasteelpark Arenberg 23, 3001, Leuven, Belgium
[12] Department of Biosystems, KU Leuven, Willem de Croylaan 42, 3001, Leuven, Belgium
[13] Department of Microbiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA
[14] Integrated Research Facility at Fort Detrick, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Fort Detrick, Frederick, MD 21702, USA
[15] Institut Pasteur, Université de Paris, Archaeal Virology Unit, F-75015, Paris, France
[16] College of Science and Engineering, Flinders University, Bedford Park, SA 5042, Australia
[17] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda MD 20894, USA
[18] Quadram Institute Bioscience, Norwich Research Park, NR4 7UQ, Norwich, UK
[19] Departments of Microbiology and Civil, Environmental, and Geodetic Engineering, Ohio State University, Columbus, OH, USA

Corresponding author: Dutilh, Bas E (bedutilh@gmail.com)

Current Opinion in Virology 2021, **51**:207–215

This review comes from a themed issue on **Virus bioinformatics**

Edited by **Alexander Gorbalenya** and **Maria Anisimova**

For complete overview about the section, refer "Virus bioinformatics (2021)"

Available online 12th November 2021

https://doi.org/10.1016/j.coviro.2021.10.011

Introduction

Few viruses are readily cultivated under laboratory conditions and even fewer cause noticeable outbreaks of disease. Over a century of virus research has resulted in an extremely biased view of global virus diversity and a limited, patchy, and non-systematic picture of the taxonomy of viruses, but viral metagenomic datasets can illuminate the true extent of the virosphere. Specifically, when previously described viral genome sequences are included in a clustering analysis together with the viral sequences obtained from metagenomics, the known sequences tend to fall within a limited subset of the clusters [1–6]. In less than a decade, the analysis of genomic sequences of uncultivated viruses, mostly derived from metagenomes, has led to a surge in virus discovery, providing invaluable new data with limited

bias for the identification and characterization of viruses. For cellular organisms (such as *Bacteria*, *Archaea*, and microbial eukaryotes), a vast expansion in the number of available genome sequences and updated analytics enabled a systematic genome-based classification that has had a profound impact on taxonomy [7•,8,9]. Techniques for cultivation-independent discovery of viruses, including metagenomic analysis and single-particle sequencing, as well as database mining efforts have contributed sequence data for hundreds of thousands of previously unknown viruses [2,10–14], with advances in overall virus taxonomy following closely behind [15,16•,17]. As genome sequences of uncultivated viruses provide new detailed information about the complex virosphere, the International Committee on Taxonomy of Viruses (ICTV) has the challenge of robustly classifying this unprecedented diversity. The ICTV has started addressing this challenge through several important amendments to policy. First, genomic characteristics are now acknowledged as the fundamental component of taxonomic classification [17], facilitating the alignment of taxonomy with the evolutionary events from which viral lineages emerged. Second, distant relationships can now be formalized using the recently introduced 15 hierarchical ranks [16•]. Although virus taxonomy will always remain dynamic to accommodate newly discovered viral lineages and adjust to advancing insights, inclusion of uncultivated viruses has already improved the accuracy and depth of the depicted evolutionary relationships of taxa and our understanding of the viruses they represent, as illustrated with several examples below [18–20]. Herein, the Bioinformatics Expert Group (BEG) of the ICTV provides a perspective on the lessons learned and remaining challenges for taxonomic classification based on viruses discovered using cultivation-free methods.

Improving taxonomic classification through computational analyses

A recurrent theme in accomplishing a robust computational framework for virus taxonomy is the power of sequence-similarity searches for identifying uncultivated viral sequences, assessing the validity and completeness of the recovered genomes, and identifying and functionally annotating genes and encoded proteins. Each newly identified virus sequence improves the potential of sequence-similarity search strategies, leading to further discovery and continually expanding our view and understanding of the virosphere. As a result of pioneering work in viruses, almost 50 years of data has been collected on genomic sequences from cultivated [21] and uncultivated [22] viruses, as well as cellular organisms in the tree of life. Computational analyses to compare and organize these data include sequence-based and profile-based searches, phylogenetic and phylogenomic tools, and clustering methods used to meaningfully identify and classify viruses in taxa at ranks from species to realm. Other tools exploit this information to distinguish viral and cellular sequences in whole-community datasets, picking out the viral needles from the metagenomic haystack [23]. Different viral strains or subtypes can be distinguished through viromics, opening up new possibilities to distinguish evolutionary and ecological dynamics of uncultivated viruses in infected hosts and in natural biomes. Recent benchmarking studies based on simulated or mock community data provide information on the advantages and disadvantages of different computational tools for identifying and classifying viruses [24–26]. Viruses that are relatively closely related to known ones can be identified by direct sequence-similarity searches of whole genomes or taxon-specific hallmark genes, which are also used for meaningful phylogenies and taxonomic classification. Viruses that are relatively unknown, representing new members within higher ranks require sensitive profile hidden Markov model-based sequence-similarity searches and assessment of the statistical significance of the hits [26]. Moreover, fundamentally different approaches have been used, including using the absence of known gene families as a signal for identifying viral sequences [27] and homology-independent features, such as genomic coding structure including directionality of genes, intergenic regions, or replichores [28,29], and nucleotide usage patterns [30]. These genomic features may be extracted computationally from viral sequences and encoded into machine-learning tools to identify viruses in metagenomic data.

To facilitate taxonomic efforts, newly discovered viral sequences need to be consistently deposited into databases. Relevant information to be recorded varies widely for different groups of viruses, depending on the extent to which they have been sampled and studied in detail. On the one hand, relevant information in highly sampled clades (such as, severe acute respiratory syndrome coronavirus 2 [SARS-CoV-2], human immunodeficiency virus 1 [HIV-1], and influenza A viruses) includes well-annotated genomic variants with detailed functional and host information. On the other hand, information on viruses from sparsely sampled taxa, which may be found in more or less exotic hosts and environments, may remain limited to non-redundant sequence clusters based on protein-sharing networks, which may be used to delineate future viral taxa [5,6,31]. Because of their versatility and scalability, gene-sharing networks have been popular for presenting preliminary taxonomic classification of viruses discovered in metagenomes.

Recent breakthroughs in high-throughput discovery of uncultivated viruses notwithstanding, there is a bottleneck in annotation and taxonomic classification. The ICTV ratifies hundreds of taxonomy proposals each year [18–20], but the rate of virus discovery is several orders of magnitude higher. As a result, the number of viruses that are represented and classified in the International Nucleotide Sequence Database Collaboration (INSDC) and

RefSeq databases, which implement ICTV's taxonomy, remains relatively limited. Databases that gather the sequences of uncultivated viruses, such as the Integrated Microbial Genome/Viral Resources (IMG/VR) database, are more inclusive but necessarily also less curated, lacking manual annotation and potentially containing occasional 'false-positives' (*i.e.*, sequences from non-viral organisms), which are an inevitable result of using computational tools to make sense of the data [32,33]. It is also important to note that the performance of virus-identification tools is often assessed in cross-validation tests, assembled by randomly extracting training and testing data from the available sequences in the database (*e.g.*, in an 80:20 ratio, respectively). Such practices should be carefully designed to account for the biased composition of databases. The sequences extracted from a database do not typically have the same degree of novelty and diversity as a real dataset, and this biased representation may result in overestimation of a tool's performance. A promising approach is to omit sequences from entire higher-rank taxa, such as viral families from the training data to mimic their novel discovery, as was done for bacterial taxa to benchmark the taxonomic classification tool CAT [34]. Such an approach depends on clear and reproducible descriptions of viral taxa, ideally according to Findability, Accessibility, Interoperability, and Reusability (FAIR) Data Principles [35]. Attaining such descriptions and making them accessible to large-scale analyses poses a major challenge for the ICTV and the international virology community.

Completing virus taxonomy with uncultivated viruses

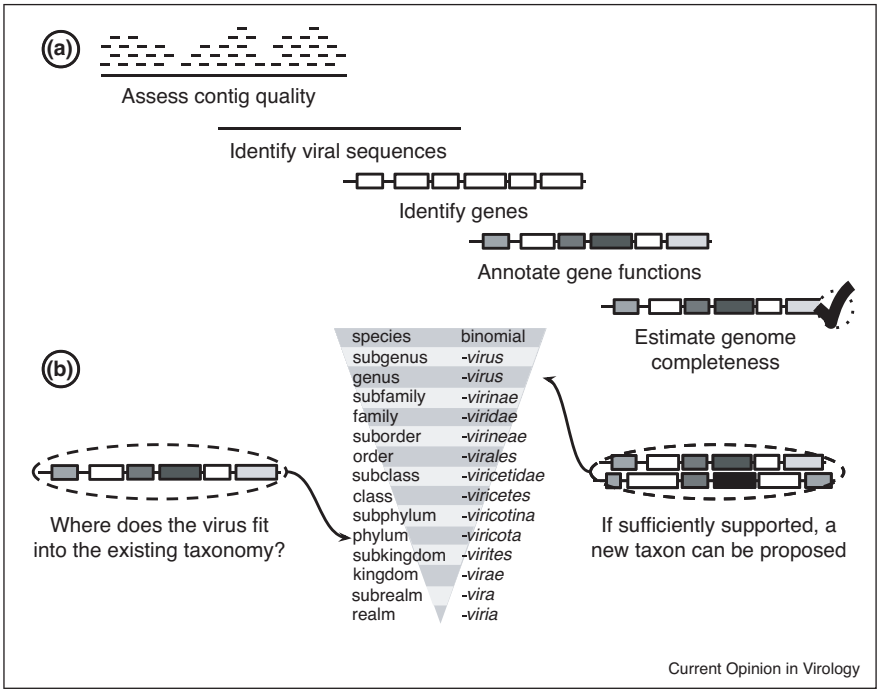
Our view of virus taxonomy is expanded by taxa that are based on viruses discovered using cultivation-free methods. Access to their genome sequences facilitates the identification of taxon-specific characteristics, including genomic properties, environmental affiliations, and biome-specific or regionally specific genomic features that may point towards host-differentiation. If virus taxonomy captures consistently evolving characteristics to represent lineages, uncultivated viruses may thus contribute new knowledge on the processes leading to the emergence of different taxa, facilitating their demarcation. As more viruses are discovered, classification becomes increasingly important to identify their relative positions in the taxonomic hierarchy (Figure 1).

When a novel virus is discovered, researchers should consider submitting a taxonomy proposal to establish a new taxon, if sufficient supporting information is available. Changes to virus taxonomy should be submitted through taxonomy proposals to the ICTV, which centralizes the process while considering the opinions of the global virology community. Consistent with the replication of a virus in nature, the repeated observation of multiple similar genomes in independent experiments is considered to be strong evidence that a sequence represents a real virus—although this is not required for an ICTV taxonomy proposal, and other evidence may be used as well. If a newly discovered virus belongs to a previously established taxon that falls within the scope of one of the ICTV Study Groups, that Study Group should be consulted and can assist in the submission of a taxonomy proposal. Taxonomy proposals are processed as part of a yearly cycle, with a deadline around May or June for submission to the relevant ICTV Subcommittee Chair, ICTV Executive Committee approvals over the summer, proposal revisions, voting by ICTV Members in October or November, and ratification around February. After ratification, the new taxonomy may be incorporated into public databases. For further details and contact information, please see the ICTV website at https://ictv.global.

Assembling and validating uncultivated virus genome sequences

Viral sequence assembly is performed with specialized software that generates contiguous sequences from short-sequencing reads, such as, Metaviral SPAdes (which exploits the specific coverage profile of viral contigs in metagenomes) [36] or SAVAGE (which can differentiate viral strains with sufficient coverage) [37]. Especially in metagenomes, sequences might be mis-assembled, so it is important to assess sequence validity. Benchmarking studies have shown that chimeras and mis-assemblies are rare but depend on the assembly program and parameter settings [38,39•]. Because chimeric assemblies may occur more frequently among less-abundant genomes with incomplete horizontal coverage, high numbers of chimeras have been observed among shorter contigs, also a consequence of low abundance [40]. Moreover, chimeras are more likely to occur among genomes of organisms that share high levels of sequence similarity, such as closely related viral populations. Another issue that might occur is artificial replication of regions or the entire contig in the assembly. Assembled sequences should be investigated for potential assembly errors—for example, by mapping the sequencing reads to the assembled contigs, observing the depth-of-coverage profile, mapping of mate-paired reads, and multi-mapping reads. Long-read sequencing technologies may be used to validate contigs assembled from short-sequencing reads or bypass short-read sequencing altogether [41]. Recovering the same or highly similar sequences multiple times independently from different datasets or samples provides strong evidence for their validity. For example, two almost-identical sequences (one polymorphism in 96 908 nucleotides) of crAss-like bacteriophages were recovered from two gorilla feces samples [42]. Validity investigations and all experimental and computational methods used to identify uncultivated virus sequences should be clearly described in the taxonomy proposal. Moreover, it is critical that the raw data are deposited in one of the

Figure 1



(a) Before taxonomic classification of uncultivated viruses, important steps that may be integrated or jointly assessed include: assessing contig/genome quality, identifying viral sequences, gene calling and functional annotation, and estimating genome completeness. **(b)** The challenge of taxonomic classification consists of: (left) placing the uncultivated viruses into the existing taxonomy at the appropriate lower rank(s) and (right) proposing new taxa at a higher rank if it is sufficiently supported by data—for example, if two or more genome sequences representing the new taxon have been observed independently.

INSDC sequence-read archives so researchers can re-examine the original data, should any doubts arise.

Estimating the completeness of uncultivated viral genome sequences is a challenge that deserves special attention, since new taxa should only be proposed if at least coding-complete genome sequences of representative viruses have been determined. Arguably the strongest indication of a genome sequence being complete is its similarity in gene content to genomes from known viruses or to sequences that were independently assembled. Features, such as terminal repeats at the ends of an assembled contig, also indicate that the genome was completely assembled and likely reflect a circular or circularly permuted genome. However, it should be noted that these repeats might represent a repeated region within the same genome. Depending on the length and identity of the repeated region, these regions could trigger assembly programs to break the genome into fragments, and the repeated regions could also end up as identical ends of the contig. The computational tool CheckV estimates genome completeness and contamination for a viral sequence by taking into consideration contig circularity and similarity of the candidate sequence to related viral genomes [43•]. Because of its dependency on reference genomes, CheckV performs less well with viral genomes that have few or distant known relatives and when terminal repeats are lacking from the contig.

Assigning viruses to the existing taxonomy: how 'novel' is a virus?

As viruses are extremely diverse, many different approaches are used for the classification of viruses into different taxa [44•]. Although several automated approaches assist in the process [6,45,46•], no single tool is capable of correctly classifying viruses of all taxa or across all ranks. Thus, the challenge of automated taxonomic classification by placing viruses into the existing taxonomy based on their sequence remains unsolved (Figure 1b). Ideally, uncultivated virus sequences would be assessed for inclusion into all existing virus taxa at all ranks using their respective demarcation criteria, which are available in the ICTV taxonomy proposals and in the yearly reports. Observing a few arbitrary genera (Figure 2) illustrates that these criteria are diverse and, in some cases, non-specific, non-concrete, and impractical. In many cases, these taxonomic demarcation criteria are the result of careful investigations into the genetic characteristics of cultured viruses that best correspond to meaningful phenotypic properties and associations with

Figure 2

Supplementary Table 9. Current species demarcation criteria from ICTV 9th and 10th reports.

Group	Family	Genus	Demarcation	Reference
dsDNA virus	Ascoviridae	Ascovirus	Phylogenetic position of genes encoding homologs of IIV6 ORFs 022L, 037L, 067R, 075L, 142R, 176R, 295L, 347L, 393L and 428L. Presence or absence of occlusion bodies Lack of DNA/DNA hybridization with other species at low stringency Restriction enzyme fragment length polymorphisms (RFLPs) Host of isolation and experimental host range Tissue tropism Association with specific hymenopteran parasites, if apparent	ICTV 10th report
		Megalocytivirus	Megalocytiviruses are distinguished from ranaviruses and lymphocystiviruses by the presence of inclusion body-bearing cells and sequence analysis of key viral genes, e.g., ATPase and MCP, for which PCR primers have been developed. Most megalocytiviruses show >94% sequence identity within these genes, whereas sequence identity with ranaviruses and lymphocystiviruses is <50%. Based on sequence analysis and serological studies, all megalocytiviruses isolated to date appear to be strains of the same or a small number of closely-related viral species. Sequence analysis suggests the presence of three closely-related clusters composed of RSIV, ISKNV, and TRBIV and a fourth, more distant, cluster comprised of a single isolate, SDDV. Whether these clusters represent distinct species, or strains of a single species, remains to be resolved. In general ISKNV-like viruses have been isolated from freshwater fish,	ICTV 10th report
		Ranavirus	Ranavirus species are distinguished by multiple criteria including amino acid and nucleotide sequence identity/similarity, phylogeny, principal host species, genome size, genetic co-linearity, gene content, and G+C content. Many isolates within the genus show >90% sequence identity/similarity within the major capsid protein and other conserved proteins. In view of this high level of sequence identity, a re-evaluation of the number of ranavirus species is currently under consideration.	ICTV 10th report
		Chloriridovirus	Only very limited colinearity has been observed between IIV3 and the genome of any other IIVs sequenced to date. The genes of IIV3, like those of other members of the family, are likely not grouped by temporal class, lack introns, are closely-spaced, and are not present on overlapping strands of the viral genome. Because suitable in vitro replication systems are lacking, little is known about the viral replication strategy. However, as with other members of the family, overall replication strategy is thought to be similar to that of FV3.	ICTV 10th report
		Iridovirus	The MCP of IIV1 shows 66.4% amino acid (aa) sequence identity to that of IIV6 and approximately 50% or lower aa sequence identity to iridovirids in other genera. Less than 1% DNA–DNA hybridization was detected by the dot-blot method between IIV1 and IIV6 genomic DNA (stringency: 26% mismatch). Restriction endonuclease profiles (HindII, EcoRI, SalI) showed a coefficient of similarity of <66% between IIV1 and IIV6. Moreover, these species did not share common antigens when tested by tube precipitation, infectivity neutralization, reversed single radial immunodiffusion or enzyme-linked immunosorbent assay. Given the current ease of sequence determination, future demarcation of viral species will likely rely more on genomic sequence analysis, host range, clinical features, etc., and less on restriction endonuclease profiles, hybridization data, and immunological cross-reactivity.	ICTV 10th report

Current Opinion in Virology

A 2019 perspective [47•] outlined Minimum Information about an Uncultivated Virus Genome (MIUViG) standards for reporting sequences of uncultivated virus genomes, including best practices and standing challenges for aspects, ranging from checking sequence validity to host prediction and abundance estimation in samples. This screenshot of the top of Supplementary Table 9 from this article [47•] lists the taxonomic demarcation criteria extracted from the International Committee on Taxonomy of Viruses (ICTV) 9th Report and 10th Report.

hosts. However, the unformalized nature of the taxon descriptions causes poor reproducibility, making it difficult for researchers to reliably assign uncultivated viruses to established taxa, even if they have been ratified by the ICTV. Although this problem is aggravated in large-scale metagenomic studies, in which thousands of sequences need to be classified at once, the scale also invites opportunity to develop systematic approaches applicable to many viral genome sequences at once, including viruses that are not (yet) covered by ICTV taxon-specific Study Groups.

The diversity in taxonomic criteria is a consequence of the diversity of viruses, their multiple evolutionary origins [48], and the diverse community of researchers involved in defining these criteria over the past 50 years [15]. While the species concept remains under debate in virology even more so than in microbiology and mycology [49–52], factors that play a role in viral evolution include the nucleic acid type, genome length, host and vector diversity, and host defense systems. A flat 95% genome-wide sequence identity threshold for species demarcation has been nearly universally adopted across bacterial and archaeal viruses, mitigating the issue of unbalanced demarcation thresholds for different virus clades [53–55]. However, demarcation criteria remain variable among eukaryotic viruses, especially at low taxonomic ranks (species and genera). For example, for uncultivated single-stranded DNA (ssDNA) viruses from the *Genomoviridae* and *Smacoviridae* families, 76–77% genome-wide pairwise identity of member viruses was chosen as a species demarcation threshold [56,57]. The reasons for this variability are rooted in the legitimate differences in the evolutionary rates, genome architectures, and replication strategies of viruses across different taxa, as well as variability in the taxonomically informative regions in viral genomes [58].

There are many ways to use virus genome sequence data for taxonomic classification. Ideally, taxonomic demarcation is based on the independent assessment of multiple genome properties with congruent conclusions. For example, the recent establishment of *Herelleviridae*, a new family of tailed bacteriophages, was based on a wide range of genomic taxonomy statistics, including marker-gene phylogenies, gene-sharing networks, and consistency in the overall genomic architecture [44•]. Different levels of genomic similarity are required for classification at different taxonomic ranks. An example comes from the analysis of samples from patients with a febrile respiratory illness from whom two papillomavirus sequences were recovered. One of the two sequences was 99.8% identical to the previously identified genome of betapapillomavirus HPV49, whereas the other clustered with a

bootstrap value of 100% among gammapapillomaviruses yet was only 61.1% identical to the closest known genome sequence [59]. These findings suggest that both a member of an established species and a new species belonging to an established genus were discovered. The availability of many closely related virus genome sequences enables the structure of the taxonomy to be resolved in much more detail than is possible with only a few viruses. For example, fine-grained typing is possible in highly sampled clades of human-infecting viruses [60], whereas gene-sharing networks are revealing a coarse-grained structure of the taxonomy of bacteriophages [6].

A minimal requirement for valid automation of virus taxonomic assignment is the availability of the genome sequences of all previously identified viruses that are part of the taxonomy. Indeed, sequence similarity is one of the strongest signals for identifying viral sequences, estimating genome completeness, identifying genes, and predicting gene functions. Thus, expanding the virus sequence database is essential to make sense of the global virosphere [61,62]. Exemplar virus genome sequences of all ICTV-ratified virus species are available in INSDC databases, a requirement for their recognition by the ICTV and a guarantee that they can be sustainably accessed. The accession numbers for these sequences are available through the virus metadata resource (VMR, see https://ictv.global/taxonomy/vmr/). However, many more virus sequences belonging to these species are present in databases, and it may be difficult to identify them without performing specific searches and in-depth sequence analyses. Ultimately, the inclusion and demarcation criteria for all taxa should be made available in a machine-readable format so that they may be programmatically accessed and readily applied to viral sequences to support their classification.

Defining genomic taxonomy for uncultivated viruses and cellular organisms

Many different approaches have been used to define virus taxa based on genomic properties. Uncultivated virus genome sequences have been clustered into approximate species-rank clusters by direct DNA–DNA sequence comparison. Thresholds of 95% average nucleotide identity over 85% of the shorter sequence length have been suggested for double-stranded DNA (dsDNA) bacterial and archaeal viruses [47•]. Clusters based on these criteria may be referred to as virus operational taxonomic units (vOTUs) and were shown to be consistent with biological species, for example in marine *Pseudoalteromonas* bacteriophages [63]. However, widely different and family specific identity thresholds are used for species demarcation of eukaryotic viruses. For instance, for uncultivated ssDNA viruses, 69–78% average nucleotide identity is used for species demarcation [64–66]. Uncultivated virus genome sequences have been clustered into approximate genus-level clusters by identifying statistically significant overlap in encoded protein content [67,68]. The clustering can take multiple forms (e.g. hierarchical clustering after pairwise comparison of genomes, clustering of gene-sharing networks with fixed or variable thresholds, or application of an empirical threshold on shared gene content). Parameters may be adjusted to tweak the cluster size, and therewith determine whether the clusters reflect slightly higher or lower taxonomic ranks.

The gold standard in genomic taxonomy of cellular organisms involves identifying widely shared marker genes and generating phylogenies of (concatenated) alignments of the encoded proteins [69,70]. In contrast to the two approaches above, this is based on the phylogeny of one or several genes, which is taken to reflect the evolutionary history of the genomes where they are found. This approach yields a highly resolved phylogenetic tree wherein taxa may be defined at multiple ranks. In the case of viruses, this approach is limited to groups that share a marker or hallmark gene. Whereas cellular organisms often share tens to hundreds of genes, even when they belong to different taxonomic domains, not a single gene is shared across all viruses. Pragmatically, this limits any taxonomic approaches based on hallmark genes to groups that share such a gene [71]. Examples include: (a) The megataxonomy of all RNA viruses in the *Riboviria* realm is based on the presence of an RNA-directed RNA polymerase (RdRp) gene [72,73]; (b) Multiple families of eukaryotic ssDNA viruses are classified within the phylum *Cressdnaviricota* based on the phylogeny of the rolling circle replication initiation endonuclease [67]; (c) The terminase large subunit TerL, HK97-like major capsid protein, and portal protein are conserved across all bacterial and archaeal *Caudoviricetes* (tailed dsDNA viruses) and eukaryotic *Herpesvirales*, and have been used as the basis to establish the realm *Duplodnaviria* [72]. Single gene/protein phylogenies might be a realistic strategy to classify viruses with small RNA and ssDNA genomes. For viruses with larger genomes such as tailed bacterial and archaeal viruses (class *Caudoviricetes*), concatenated protein phylogenies become practical [46•]. Critically, the phylogenies of individual marker genes should be compared to a phylogeny based on a concatenated alignment to assess potential horizontal gene transfer and taxonomic biases [74–76].

Conclusion: aligning computational classifications with ICTV-ratified taxa

Bioinformaticians have created computational tools for genomic taxonomy that cluster viruses based on genome similarity [6,45,77–80]. Although these tools provide a valuable first-order estimate of virus taxa, especially at lower ranks, they rarely assess hierarchical taxonomic structure across all ranks and may conflict with ICTV-ratified taxa that have been meticulously defined by experts. The main reason for this discrepancy is the fact that most current computational tools are based on a

single genomic character, whereas demarcation criteria for ICTV-ratified taxa are variable (Figure 2). By formulating taxonomic inclusion and demarcation criteria in a specific and concrete manner, bioinformaticians can help disclose virus taxonomy and assist the ICTV in incorporating the diverse demarcation criteria into computational tools and models. A notable collaboration to develop sequence-based taxonomy of cellular organisms (*Bacteria* and *Archaea*) resulted in the prokaryotic Genome Taxonomy Database (GTDB), which includes metagenome-assembled genomes, and has led to significant community engagement [81,82]. The open call for taxonomy proposals by the ICTV enables all virologists to participate and contribute to charting the structure of the virosphere.

Conflict of interest statement

Nothing declared.

Acknowledgements

The authors thank Anya Crane (National Institutes of Health [NIH] National Institute of Allergy and Infectious Diseases [NIAID]) for critically editing the manuscript. The authors are members of the ICTV Virus Bioinformatics Expert Group.

BED is supported by the European Research Council (ERC) Consolidator grant 865694: DiversiPHI. CL is supported by the Research Foundation Flanders (FWO) SB grant 1S64720N. RAE is supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (NIH) under Award Number RC2DK116713. SR is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was also supported in part through Laulima Government Solutions, LLC, prime contract with the NIH National Institute of Allergy and Infectious Diseases (NIAID) under Contract No. HHSN272201800013C. JHK performed this work as an employee of Tunnell Government Services (TGS), a subcontractor of Laulima Government Solutions, LLC, under Contract No. HHSN272201800013C. SS acknowledges partial support from the Special Research Initiative (MAFES), Mississippi State University, and the National Institute of Food and Agriculture, U.S. Department of Agriculture, Hatch Project 1021494. EMA acknowledges the support of the Biotechnology and Biological Sciences Research Council (BBSRC); this research was funded by the BBSRC Institute Strategic Programme Gut Microbes and HealthBB/R012490/1 and its constituent projects BBS/E/F/000PR10353 and BBS/E/F/000PR10356. This research was supported in part by the Intramural Research Program of the National Library of Medicine at the NIH, National Library of Medicine. MBS was supported by the U.S. National Science Foundation award ABI#1759874.

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results. The views and conclusions contained in this publication are those of the authors. The content of this publication does not necessarily reflect the views or policies, either expressed or implied, of the US Department of Health and Human Services (HHS) or of the institutions and companies affiliated with the authors. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest

1. de Jonge PA, von Meijenfeldt FAB, Costa AR, Nobrega FL, Brouns SJJ, Dutilh BE: **Adsorption sequencing as a rapid method to link environmental bacteriophages to hosts**. *iScience* 2020, **23**:101439.

2. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB: **The gut virome database reveals age-dependent patterns of virome diversity in the human gut**. *Cell Host Microbe* 2020, **28**:724-740.e728.

3. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, Singleton CM, Solden LM, Naas AE, Boyd JA *et al.*: **Host-linked soil viral ecology along a permafrost thaw gradient**. *Nat Microbiol* 2018, **3**:870-880.

4. Gregory AC, Zayed AA, Conceicao-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C *et al.*: **Marine DNA viral macro- and microdiversity from pole to pole**. *Cell* 2019, **177**:1109-1123 e1114.

5. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E, Poulain J *et al.*: **Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses**. *Nature* 2016, **537**:689-693.

6. Jang HB, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R *et al.*: **Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks**. *Nat Biotechnol* 2019, **37**:632-639.

7. • Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P: **A complete domain-to-species taxonomy for bacteria and archaea**. *Nat Biotechnol* 2020, **38**:1079-1086
Standardizes prokaryotic taxonomy based on phylogenetic analysis of shared marker genes with high community involvement.

8. Rinke C, Chuvochina M, Mussig AJ, Chaumeil P-A, Davín AA, Waite DW, Whitman WB, Parks DH, Hugenholtz P: **A standardized archaeal taxonomy for the genome taxonomy database**. *Nat Microbiol* 2021, **6**:946-959.

9. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K *et al.*: **A new view of the tree of life**. *Nat Microbiol* 2016, **1**:16048.

10. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD: **Massive expansion of human gut bacteriophage diversity**. *Cell* 2021, **184**:1098-1109.e1099.

11. Dávila-Ramos S, Castelán-Sánchez HG, Martínez-Ávila L, Sánchez-Carbente MDR, Peralta R, Hernández-Mendoza A, Dobson ADW, Gonzalez RA, Pastor N, Batista-García RA: **A review on viral metagenomics in extreme environments**. *Front Microbiol* 2019, **10**:2403.

12. Kristensen DM, Mushegian AR, Dolja VV, Koonin EV: **New dimensions of the virus world discovered through metagenomics**. *Trends Microbiol* 2010, **18**:11-19.

13. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, Proal AD, Fischbach MA, Bhatt AS, Hugenholtz P *et al.*: **Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome**. *Nat Microbiol* 2021, **6**:960-970.

14. Obbard DJ: **Expansion of the metazoan virosphere: progress, pitfalls, and prospects**. *Curr Opin Virol* 2018, **31**:17-23.

15. Adams MJ, Lefkowitz EJ, King AMQ, Harrach B, Harrison RL, Knowles NJ, Kropinski AM, Krupovic M, Kuhn JH, Mushegian AR *et al.*: **50 years of the International Committee on Taxonomy of Viruses: progress and prospects**. *Arch Virol* 2017, **162**:1441-1446.

16. • Gorbalenya AE, Krupovic M, Mushegian AR, Kropinski AM, Siddell SG, Varsani A, Adams MJ, Davison AJ, Dutilh BE, Harrach B *et al.*: **The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks**. *Nat Microbiol* 2020, **5**:668-674
Presents fifteen-rank virus taxonomy, opens new possibilities for classifying distinct viruses.

17. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B *et al.*: **Consensus statement: virus taxonomy in the age of metagenomics**. *Nat Rev Microbiol* 2017, **15**:161-168.

18. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Alfenas-Zerbini P, Davison AJ, Dempsey DM, Dutilh BE, Garcia ML *et al.*: **Changes to virus taxonomy and to the international code of virus classification and**

nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Arch Virol* 2021, **166**:2633-2648.

19. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Dempsey DM, Dutilh BE, Harrach B, Harrison RL, Hendrickson RC *et al.*: **Changes to virus taxonomy and the statutes ratified by the International Committee on Taxonomy of Viruses (2020)**. *Arch Virol* 2020, **165**:2737-2748.

20. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Dempsey DM, Dutilh BE, Harrach B, Harrison RL, Hendrickson RC, Junglen S *et al.*: **Changes to virus taxonomy and the international code of virus classification and nomenclature ratified by the International Committee on Taxonomy of Viruses (2019)**. *Arch Virol* 2019, **164**:2417-2429.

21. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: **Nucleotide sequence of bacteriophage φX174 DNA**. *Nature* 1977, **265**:687-695.

22. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F: **Genomic analysis of uncultured marine viral communities**. *Proc Natl Acad Sci U S A* 2002, **99**:14250-14255.

23. Soueidan H, Schmitt L-A, Candresse T, Nikolski M: **Finding and identifying the viral needle in the metagenomic haystack: trends and challenges**. *Front Microbiol* 2014, **5**:739.

24. Glickman C, Hendrix J, Strong M: **Simulation study and comparative evaluation of viral contiguous sequence identification tools**. *BMC Bioinformatics* 2021, **22**:329.

25. Ho SFS, Millard AD, van Schaik W: **Comprehensive benchmarking of tools to identify phages in metagenomic shotgun sequencing data**. *bioRxiv* 2021 http://dx.doi.org/10.1101/2021.04.12.438782v1.

26. Pratama AA, Bolduc B, Zayed AA, Zhong Z-P, Guo J, Vik DR, Gazitúa MC, Wainaina JM, Roux S, Sullivan MB: **Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation**. *PeerJ* 2021, **9**:e11447.

27. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC: **Uncovering earth's virome**. *Nature* 2016, **536**:425-430.

28. Akhter S, Aziz RK, Edwards RA: **PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies**. *Nucleic Acids Res* 2012, **40**:e126.

29. Roux S, Enault F, Hurwitz BL, Sullivan MB: **VirSorter: mining viral signal from microbial genomic data**. *PeerJ* 2015, **3**:e985.

30. Willner D, Thurber RV, Rohwer F: **Metagenomic signatures of 86 microbial and viral metagenomes**. *Environ Microbiol* 2009, **11**:1752-1766.

31. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R: **Reticulate representation of evolutionary and functional relationships between phage genomes**. *Mol Biol Evol* 2008, **25**:762-777.

32. Paez-Espino D, Roux S, Chen IMA, Palaniappan K, Ratner A, Chu K, Huntemann M, Reddy TBK, Pons JC, Llabrés M *et al.*: **IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes**. *Nucleic Acids Res* 2019, **47**:D678-D686.

33. Roux S, Páez-Espino D, Chen IMA, Palaniappan K, Ratner A, Chu K, Reddy TBK, Nayfach S, Schulz F, Call L *et al.*: **IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses**. *Nucleic Acids Res* 2021, **49**:D764-D775.

34. von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE: **Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT**. *Genome Biol* 2019, **20**:217.

35. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE *et al.*: **The FAIR guiding principles for scientific data management and stewardship**. *Sci Data* 2016, **3**:160018.

36. Antipov D, Raiko M, Lapidus A, Pevzner PA: **Metaviral SPAdes: assembly of viruses from metagenomic data**. *Bioinformatics* 2020, **36**:4126-4129.

37. Baaijens JA, Aabidine AZE, Rivals E, Schonhuth A: **De novo assembly of viral quasispecies using overlap graphs**. *Genome Res* 2017, **27**:835-848.

38. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB: **Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity**. *PeerJ* 2017, **5**:e3817.

39. • Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E *et al.*: **Critical assessment of metagenome interpretation-a benchmark of metagenomics software**. *Nat Methods* 2017, **14**:1063-1071
Independently benchmarks metagenome interpretation software with high community involvement.

40. Mende DR, Waller AS, Sunagawa S, Järvelin AI, Chan MM, Arumugam M, Raes J, Bork P: **Assessment of metagenomic assembly using simulated next generation sequencing data**. *PLoS One* 2012, **7**:e31386.

41. Latorre-Pérez A, Villalba-Bermell P, Pascual J, Vilanova C: **Assembly methods for nanopore-based metagenomic sequencing: a comparative study**. *Sci Rep* 2020, **10**:13588.

42. Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, Dinsdale EA, Cinek O, Aziz RK, McNair K, Barr JJ *et al.*: **Global phylogeography and ancient evolution of the widespread human gut virus crAssphage**. *Nat Microbiol* 2019, **4**:1727-1736.

43. • Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC: **CheckV assesses the quality and completeness of metagenome-assembled viral genomes**. *Nat Biotechnol* 2021, **39**:578-585
Quantitatively estimates virus genome sequence completeness.

44. • Barylski J, Enault F, Dutilh BE, Schuller MB, Edwards RA, Gillis A, Klumpp J, Knezevic P, Krupovic M, Kuhn JH *et al.*: **Analysis of spounaviruses as a case study for the overdue reclassification of tailed phages**. *Syst Biol* 2020, **69**:110-123
Presents new caudovirus taxonomy based on complementary lines of genomics evidence.

45. 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.

46. • Low SJ, Džunková M, Chaumeil P-A, 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
Presents new caudovirus taxonomy using an approach analogous to Ref. [7•].

47. • Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A *et al.*: **Minimum Information about an Uncultivated Virus Genome (MIUViG)**. *Nat Biotechnol* 2019, **37**:29-37
Outlines best practices for reporting genome sequences of uncultivated viruses.

48. Krupovic M, Dolja VV, Koonin EV: **Origin of viruses: primordial replicators recruiting capsids from hosts**. *Nat Rev Microbiol* 2019, **17**:449-458.

49. Bobay L-M, Ochman H: **Biological species in the viral world**. *Proc Natl Acad Sci U S A* 2018, **115**:6040-6045.

50. Peterson AT: **Defining viral species: making taxonomy useful**. *Virol J* 2014, **11**:131.

51. Van Regenmortel MHV, Ackermann H-W, Calisher CH, Dietzgen RG, Horzinek MC, Keil GM, Mahy BWJ, Martelli GP, Murphy FA, Pringle C *et al.*: **Virus species polemics: 14 senior virologists oppose a proposed change to the ICTV definition of virus species**. *Arch Virol* 2013, **158**:1115-1119.

52. Zachos FE: *Species Concepts in Biology: Historical Development*. 2016.

53. Adriaenssens EM, Krupovic M, Knezevic P, Ackermann H-W, Barylski J, Brister JR, Clokie MRC, Duffy S, Dutilh BE, Edwards RA *et al.*: **Taxonomy of prokaryotic viruses: 2016 update from the ICTV Bacterial and Archaeal Viruses Subcommittee**. *Arch Virol* 2017, **162**:1153-1157.

54. Adriaenssens EM, Sullivan MB, Knezevic P, van Zyl LJ, Sarkar BL, Dutilh BE, Alfenas-Zerbini P, Łobocka M, Tong Y, Brister JR *et al.*: **Taxonomy of prokaryotic viruses: 2018-2019 update from the ICTV Bacterial and Archaeal Viruses Subcommittee**. *Arch Virol* 2020, **165**:1253-1260.

55. Adriaenssens EM, Wittmann J, Kuhn JH, Turner D, Sullivan MB, Dutilh BE, Jang HB, van Zyl LJ, Klumpp J, Lobocka M *et al.*: **Taxonomy of prokaryotic viruses: 2017 update from the ICTV Bacterial and Archaeal Viruses Subcommittee**. *Arch Virol* 2018, **163**:1125-1129.

56. Varsani A, Krupovic M: **Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family *Genomoviridae***. *Virus Evol* 2017, **3**:vew037.

57. Varsani A, Krupovic M: ***Smacoviridae*: a new family of animal-associated single-stranded DNA viruses**. *Arch Virol* 2018, **163**:2005-2015.

58. Moreno-Gallego JL, Reyes A: **Informative regions in viral genomes**. *Viruses* 2021, **13**:1164.

59. Mokili JL, Dutilh BE, Lim YW, Schneider BS, Taylor T, Haynes MR, Metzgar D, Myers CA, Blair PJ, Nosrat B *et al.*: **Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness**. *PLoS One* 2013, **8**:e58404.

60. Nooij S, Schmitz D, Vennema H, Kroneman A, Koopmans MPG: **Overview of virus metagenomic classification methods and their biological applications**. *Front Microbiol* 2018, **9**:749.

61. Canuti M, van der Hoek L: **Virus discovery: are we scientists or genome collectors?** *Trends Microbiol* 2014, **22**:229-231.

62. Dutilh BE: **Metagenomic ventures into outer sequence space**. *Bacteriophage* 2014, **4**:e979664.

63. Duhaime MB, Solonenko N, Roux S, Verberkmoes NC, Wichels A, Sullivan MB: **Comparative omics and trait analyses of marine pseudoalteromonas phages advance the phage OTU concept**. *Front Microbiol* 2017, **8**:1241.

64. Krupovic M, Varsani A: **A 2021 taxonomy update for the family *Smacoviridae***. *Arch Virol* 2021, **166**:3245-3253.

65. Varsani A, Krupovic M: **Family *Genomoviridae*: 2021 taxonomy update**. *Arch Virol* 2021, **166**:2911-2926.

66. Varsani A, Opriessnig T, Celer V, Maggi F, Okamoto H, Blomstrom AL, Cadar D, Harrach B, Biagini P, Kraberger S: **Taxonomic update for mammalian anelloviruses (family *Anelloviridae*)**. *Arch Virol* 2021, **166**:2943-2953.

67. Krupovic M, Varsani A, Kazlauskas D, Breitbart M, Delwart E, Rosario K, Yutin N, Wolf YI, Harrach B, Zerbini FM *et al.*: ***Cressdnaviricota*: a virus phylum unifying seven families of rep-encoding viruses with single-stranded, circular DNA genomes**. *J Virol* 2020, **94**.

68. Roux S, Krupovic M, Daly RA, Borges AL, Nayfach S, Schulz F, Sharrar A, Matheus Carnevali PB, Cheng J-F, Ivanova NN *et al.*: **Cryptic inoviruses revealed as pervasive in bacteria and archaea across earth's biomes**. *Nat Microbiol* 2019, **4**:1895-1906.

69. Hugenholtz P, Chuvochina M, Oren A, Parks DH, Soo RM: **Prokaryotic taxonomy and nomenclature in the age of big sequence data**. *ISME J* 2021, **15**:1879-1892.

70. Thompson CC, Amaral GR, Campeão M, Edwards RA, Polz MF, Dutilh BE, Ussery DW, Sawabe T, Swings J, Thompson FL: **Microbial taxonomy in the post-genomic era: rebuilding from scratch?** *Arch Microbiol* 2015, **197**:359-370.

71. Turner D, Kropinski AM, Adriaenssens EM: **A roadmap for genome-based phage taxonomy**. *Viruses* 2021, **13**:506.

72. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH: **Global organization and proposed megataxonomy of the virus world**. *Microbiol Mol Biol Rev* 2020, **84**:e00061-19.

73. Gorbalenya AE, Krupovic M, Siddell SG, Varsani A, Kuhn JH: **Riboviria: establishing a single taxon that comprises RNA viruses at the basal rank of virus taxonomy**. *International Committee on Taxonomy of Viruses (ICTV)* 2017 https://talk.ictvonline.org/ictv/proposals/2017.2006G.A.v2013.Riboviria.zip.

74. Ignacio-Espinoza JC, Sullivan MB: **Phylogenomics of T4 cyanophages: lateral gene transfer in the 'core' and origins of host genes**. *Environ Microbiol* 2012, **14**:2113-2126.

75. Rossi A, Treu L, Toppo S, Zschach H, Campanaro S, Dutilh BE: **Evolutionary study of the crassphage virus at gene level**. *Viruses* 2020, **12**:1035.

76. Wennmann JT, Keilwagen J, Jehle JA: **Baculovirus Kimura two-parameter species demarcation criterion is confirmed by the distances of 38 core gene nucleotide sequences**. *J Gen Virol* 2018, **99**:1307-1320.

77. Bao Y, Chetvernin V, Tatusova T: **PAirwise Sequence Comparison (PASC) and its application in the classification of filoviruses**. *Viruses* 2012, **4**:1318-1327.

78. Lauber C, Gorbalenya AE: **Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses**. *J Virol* 2012, **86**:3890-3904.

79. Moraru C, Varsani A, Kropinski AM: **VIRIDIC-a novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses**. *Viruses* 2020, **12**:1268.

80. Muhire BM, Varsani A, Martin DP: **SDT: a virus classification tool based on pairwise sequence alignment and identity calculation**. *PLoS One* 2014, **9**:e108277.

81. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH: **GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database**. *Bioinformatics* 2020, **36** 1927–1927.

82. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P: **A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life**. *Nat Biotechnol* 2018, **36**:996-1004.