Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Virology

# Bioinformatics of virus taxonomy: foundations and tools for developing sequence-based hierarchical classification

Alexander E Gorbalenya[1,2] and Chris Lauber[3]

The genome sequence is the only characteristic readily obtainable for all known viruses, underlying the growing role of comparative genomics in organizing knowledge about viruses in a systematic evolution-aware way, known as virus taxonomy. Overseen by the International Committee on Taxonomy of Viruses (ICTV), development of virus taxonomy involves taxa demarcation at 15 ranks of a hierarchical classification, often in host-specific manner. Outside the ICTV remit, researchers assess fitting numerous unclassified viruses into the established taxa. They employ different metrics of virus clustering, basing on conserved domain(s), separation of viruses in rooted phylogenetic trees and pair-wise distance space. Computational approaches differ further in respect to methodology, number of ranks considered, sensitivity to uneven virus sampling, and visualization of results. Advancing and using computational tools will be critical for improving taxa demarcation across the virosphere and resolving rank origins in research that may also inform experimental virology.

**Addresses**
[1] Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands

[2] Faculty of Bioengineering and Bioinformatics and Belozersky, Institute of Physico-Chemical Biology, Lomonosov Moscow State University, 119899, Moscow, Russia

[3] Institute for Experimental Virology, TWINCORE Centre for Experimental and Clinical Infection Research, A Joint Venture between the Hannover Medical School (MHH) and the Helmholtz Centre for Infection Research (HZI), Hannover, Germany

Corresponding author:
Gorbalenya, Alexander E (a.e.gorbalenya@lumc.nl)

**Current Opinion in Virology** 2022, **52**:48–56

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

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

https://doi.org/10.1016/j.coviro.2021.11.003



## Introduction

Viruses are the first biological entities whose genomes were fully sequenced, and this characterization is often the only available for known viruses [1]. Being accurate and affordable, full genome sequencing revolutionized the practice of virology and gave rise to comparative genomics that forms the core of bioinformatics. Assisted by the growing number of diverse software tools, researchers were empowered to detect and rationalize nucleotide and amino acid sequence patterns that evolved under various constraints in the respective virus biopolymers [2]. These insights have advanced our understanding of the function, structure, and evolution of viruses.

Comparative genomics is also highly conductive for the development of a hierarchical classification for virus taxonomy, which additionally includes taxa nomenclatures [3,4]. This research venue was initiated more than 50 years ago, and not long before comparative genomics, with the aim of organizing a growing knowledge about viruses in a systematic way [5•]. The International Committee on Taxonomy of Viruses (ICTV), which has no counterparts elsewhere in biology with a comparable broad remit of decision-making on all matters of classification and nomenclature, oversees the development of virus taxonomy. The original hierarchical classification of virus taxonomy included just few ranks, with family and genus being most prominent and actively used. This unusually limited rank choice was owing to the only few dozen virus pathogens known at the time and linked to the extremely fast mutation rate of viruses, generating astonishing diversity of forms and phenotypes that seemed barely connected in the pre-genomic era. The advent of comparative genomics along with downstream phylogenetic analysis both helped and challenged expert ICTV Study Groups (SG) who are in charge of taxonomy development of various virus families. Results from these analyses and associated conceptualizations have gradually propelled a notion that the formal virus classification could be developed in a manner that reflects the evolution of viruses by descent; much like it was realized for taxonomies of cellular life forms many decades before. This notion gained urgency after the advent of increasingly affordable high-throughput genome sequencing (HGS) that enabled large scale genomic characterization of viruses in biological specimens (metagenomics and transcriptomics) [6–8].

Accordingly, ICTV recently went to (1) endorse comparative genomics as a sole acceptable basis for the recognition of virus taxa [5•,9•]; and (2) adapt a Linnaean-like rank structure of the formal classification of the virosphere that includes 15 ranks [10•,11,12]. Researchers

use bioinformatics methods for establishing new virus taxa at an accelerating rate in approaches that are informed by specifics of viruses which may differ from cellular organisms in one or more aspects. First of all, there is no comparable virus counterpart to the vast phenotypic-based taxonomy of cellular organisms developed over many decades, since only several hundred virus taxa had been recognized phenotypically before the taxonomy development has become tightly coupled to genomics-based discovery of viruses over the last decade [13]. Second, there is no common molecular denominator that is shared by all viruses (alike to ribosomal RNA of cellular organisms), whose variation could be evaluated for reconstruction of a virosphere-wide tree, comparable to the tree of Life (ToL) [14,15•]. Third, calibration of the time-line of virus taxa is indirect and inferred using computational genomics of virus-host associations [16,17,18], although see Ref. [19]. Finally, virologists still debate about the meaning of species that populate the most numerous and principal rank of taxonomy [20,21,22•], which is firmly population-based in other taxonomies [23,24]. These and other specifics make advancements of virus taxonomy extremely challenging and somewhat controversial, but also critical for reinforcing the connection between exploration of the virosphere and diverse viromes with experimental and applied research [3].

In this respect, the use of common software by different SGs for virus genome-based classification would be a promising way forward in the young field lacking a gold standard. It would also empower any interested party to use the developed taxonomy most efficiently in quality manner and with insight that may not be available otherwise. Below we substantiate this opinion in an overview of the background and computational approaches that have been specifically developed in the field of virus taxonomy.

## Developing versus using virus taxonomy

To learn about virus taxonomy, researchers are advised to consult the ICTV web site that provides access to (1) a hierarchical rank structure covering divergence of the entire virosphere; (2) names of recognized taxa, from species to realms; (3) a list of GenBank IDs of reference viruses, typically one virus per known species and under 10 000 totally; (4) approved taxonomical proposals that often include other classified viruses as well as demarcation criteria detailing the delimitation of taxa at the species and possibly other ranks; (5) the online 10th ICTV Report including taxonomy of virus families [10•,25]. These resources are defined by the ICTV remit; they may be sufficient for researchers who want to overview the current state of the art or advance taxonomy further.

However, the ICTV remit does not include other important aspects that are critical for the use of taxonomy by the field, namely the classification of *every* known virus and quality control mechanism for ensuring proper application of the developed taxonomy to viruses. Combined with the lack of consistency of taxa delimitation across virus taxonomy [26], they leave un-classified many thousands of viruses that are not on the ICTV list. This gap is largely addressed by the National Center for Biotechnology Information (NCBI) GenBank team who maintain web pages devoted to virus taxonomy [27]. They satisfy database standards on genome sequences of any origin, although they do it with a disclaimer that is worth of reproducing here: 'The NCBI taxonomy database is not an authoritative source for nomenclature or classification — please consult the relevant scientific literature for the most reliable information.' Notwithstanding this notable reservation, the NCBI site is effectively the only available resource for taxonomy for the majority of viruses.

## Computational virus taxonomy: approaches, software and challenges

The current computational approaches assisting virus taxonomy may be based on: (1) patterns of nucleotide or amino acid sequence variation, (2) gene or protein content by homology, (3) phylogeny, and (4) pair-wise genetic distance. The choice of metric(s) may be dictated by practical considerations of a particular study and often affected by virus host(s), although these approaches are not mutually exclusive and the conformity of an inferred classification with evolution by descent is most desirable.

The simplest approaches are based on a summary statistic of sequences. For instance, the abundance of certain combinations of nucleotides, for example, G + C content, or a count of all possible oligonucleotides or — peptides of a certain length- (K-mer methods), also in combination with word position, may be used to produce virus classifications [28,29]. However, the extent to which these metrics are sensitive to data-specific attributes such as uneven virus sampling, recombination, convergence, and lineage-specific and region-specific rates of evolution may exceed those of the approaches that are based on sequence alignments which facilitate addressing these complications (described below). Also, interpretability of the differences underlying the K-mer methods is often hardly accessible.

### Gene content and network criterion

Conservation of a gene or protein (domain) in a group of viruses defines gene content and associated gene order (synteny), known also as genome organization. These characteristics have been used in informal virus classification for decades [30–33], serving also as markers and denominators of many taxa established at different ranks by SG experts (see for instance [34] for a capsid marker of the order *Picornavirales* and [35,36] for replicase markers of the order *Nidovirales*). Recently, this approach was formalized in VConTACT [37,38•], which analyzes gene

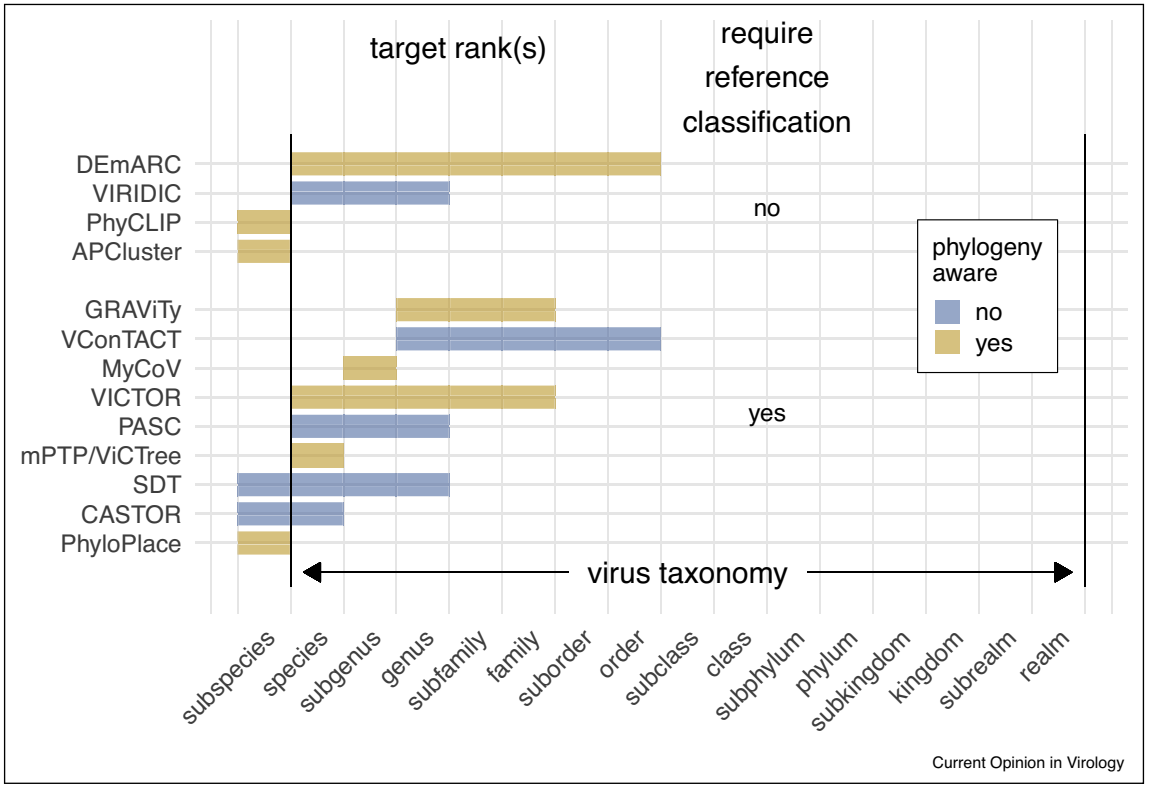
sharing networks, and GRAViTy [39,40], which analyzes gene content and genomic organization signatures that constitute a set of polythetic characteristics (Figure 1 and Table 1). An advanced version 2.0 of VConTACT enables automatic assignment of phages of large metagenomic datasets to new taxa at the genus-family ranks. Each of these tools uses different metrics and statistics to infer a virus clustering. They may work most efficiently when applied to viruses from very many families in an analysis that requires external expert-based classification of selected viruses for comparison. Upon evaluation, these tools demonstrated high sensitivity and specificity in studies of eukaryotic viruses (GRAViTy at the genus-family range) and DNA and RNA phages (VConTACT at the genus-order range), with the latter accommodating frequent gene exchange. Notably, the employed metrics may not be applied to outliers and singletons, and it remains unknown whether they may capture signals for taxa delimitation at the species rank.

These and other network-based analyses involve all proteins that are conserved in the analyzed viruses. Its core set is limited and includes capsid proteins and various polymerase proteins (RNA-directed RNA polymerase, reverse transcriptase, protein-primed family B DNA polymerases) [41]. Quantitative analysis of the relationship of these proteins builds the foundation of two other major approaches to taxonomy discussed below (see sections 'Monophyletic clusters and tree topology criterion' and 'Pair-wise distances and threshold criterion').

### Monophyletic clusters and tree topology criterion

Conserved proteins are commonly used to generate unrooted or, ideally, rooted trees, and tree topology is then analyzed to identify monophyletic clusters, which are among the major criteria in non-virus systematics [42,43] and was mandated for virus species delimitation only recently. Maximum-likelihood phylogenies, commonly computed using high-performance tools such as PhyML [44], FastTree [45], RaxML [46], and IQ-Tree [47] are standard in virus taxonomy. Bayesian phylogenies reconstructed using programs like MrBayes [48] or BEAST [49] may constitute alternatives that are at least competitive but which typically come along with higher computational costs. In the expert-based virus taxonomy framework, external information about virus phenotype

**Figure 1**




Scope and attributes of bioinformatics software for virus taxonomy.
The bars indicate the range of taxonomic ranks that is applicable to the respective software tool, according to the original publication, although some of these ranks may not have been tested practically. Note that this range of ranks does not necessarily correspond to the actual number of ranks in the output of a tool, as listed in Table 1, and characteristics of a tool may change with its advancement, for example, see VConTACT. Yellow bar color is used to indicate phylogeny-aware tools; other tools have blue bar color. Four tools that do not require a reference classification are shown on top, others at the bottom part. For other details of software see Table 1 and text.

**Table 1**

**Software tools for virus taxonomy**

| Software/method [Ref] | Strategy[a] | Input | Output | Output levels (ranks)[b] | External versus data-driven[c] | Additional notes |
|---|---|---|---|---|---|---|
| PASC [57•] | PD | New genome sequences + reference distance matrix + reference classification | Classification of *new* viruses | 2 (S,G) | E | Developed for fast web-based genome-wide analysis |
| PhyloPlace [66] | P | New and reference genome sequences + reference classification | Classification of *new* viruses | 1 (SS) | E | Developed for HCV and HIV-1 subtypes |
| APCluster [68] | P | Pairwise distance matrix of reference and new viruses | Classification of *all* viruses in data set | 1 (SS) | D | Applied to intra-species (Rabies virus) classification |
| DEmARC [59,70•] | PD, [P], C, TO | Multiple sequence alignment, distance matrix or phylogenetic tree of reference and new viruses | Classification of *all* viruses in data set | Many, user-defined | D | Developed for large monophyletic groups of viruses |
| SDT [58•] | PB | Genome or protein sequences of new and reference sequences + reference classification | Classification of *new* viruses | 3 (SS, S, G) | E | Developed for new viruses from a monophyletic group |
| VConTACT [37,38•] | PC, N | New and reference genome sequences + reference classification | Classification of *new* viruses | 1 | E | Developed for archaeal and bacterial viruses |
| mPTP / ViCTree [54,53] | P | New and reference genome sequences + reference classification | Classification of *new* viruses | 1 | E | Developed for viruses from a monophyletic group |
| VICTOR [60] | PD, P, C, TO | New and reference genome sequences + reference classification | Classification of *all* viruses in data set | 3–4 (S, G, [SF], F) | E | Developed for archaeal and prokaryotic viruses |
| CASTOR [65] | ML | New and reference genome sequence (CASTOR-predict) or genome sequences + classification (CASTOR-build) | Classification of *new* or *all* viruses | 2 (SS, S) | E | Developed to classify new viruses or build new classification |
| GRAViTy [39,40] | PC, PD, TO | New and reference genome sequences + reference classification | Classification of *new* viruses | 2 (G, F) | E | Developed for classification into existing taxa (family or higher) |
| PhyCLIP [67] | PD, P, TO | Phylogenetic tree + distance threshold offset + false discovery rate for reference and new viruses | Classification of *all* viruses in data set | 1 | D | Developed for intra-species (Influenza virus) classification |
| MyCoV [64] | PD | New and reference genome sequences + taxon identifier | Classification of *new* viruses | 1 (SG) | E | Developed for coronavirus subgenus classification |
| VIRIDIC [63] | PD, C | New and reference genome sequences + distance thresholds | Classification of *all* viruses in data set | 2 (S, G) | D | Developed for prokaryotic viruses |

[a] Based on pairwise distance (PD), phylogeny (P), clustering (C), distance threshold optimization (TO), protein content (PC), network (N) and machine learning (ML).
[b] Subspecies (SS), species (S), genus (G), subfamily (SF), family (F).
[c] External (E) versus data-driven (D) classification. The E-dependent packages may leave a sequence unclassified if it does not fit into any existing taxon.

or genotype is used to demarcate monophyletic clusters as taxa, and the shared characteristic(s) known as denominator(s) may continue serving as marker(s) of the respective taxa. Concatenated conserved protein domains may offer improved resolution for phylogeny-based taxonomy as was demonstrated for the order *Caudovirales* [50•] and used for advancement of the order *Nidovirales* [51,52•]. The phylogeny-based application to the most conserved protein domains defines taxa at recently established ranks above the order rank [15•], which are outside the scope of the available software packages (Figure 1), either by design or due to challenges with interpretation of results.

It would be desirable to seek inter-taxa consistency of demarcation criteria upon rank assignment of taxa, which in practice is rarely pursued systematically. To address this aspect, the VicTree software uses a multi-rate Poisson Tree Processes (mPTP) method to demarcate taxa at the species rank [53]. The mPTP [54] approach is part of the growing trend of developing phylogeny-aware tools to assist with species identification of cellular organisms [24]. Their use in virology is encouraging but complicated by technical issues related to relatively high rates of virus evolution. Also its application beyond the species rank may be unrealistic.

### Pair-wise distances and threshold criterion

If evolution proceeds according to a molecular clock model, taxa delimitation that is otherwise based on cumbersome analysis of tree topology may be simplified. Under this condition, pairwise genetic distances between viruses that constitute the terminal tips of phylogenetic trees become sufficient to inform clustering, without tree reconstruction (although see below) [55]. The utility of this measure for the development of virus taxonomy was recognized early on in the so-called pairwise sequence comparison analysis [56]. This approach was subsequently implemented into a dedicated tool, called PASC [57•], that is run at the NCBI web site for analysis of many virus families (https://www.ncbi.nlm.nih.gov/sutils/pasc/). This and other similar implementations, for example, the popular SDT software [58•], utilize a frequency distribution of pairwise sequence divergence, commonly measured in % identical residues between pairs of virus genomes, genes or proteins. Minima in the density distribution or external considerations, or both, inform expert decisions about positions of divergence thresholds that separate ranks of classification. These thresholds allow for the assignment of new viruses to existing taxa, as well as the demarcation of new ranks and taxa. A common criticism of this approach is the dependence of threshold assignment on the density distribution of pair-wise distances that makes it highly sensitive to data sampling, which is typically uneven and highly biased for many virus groups.

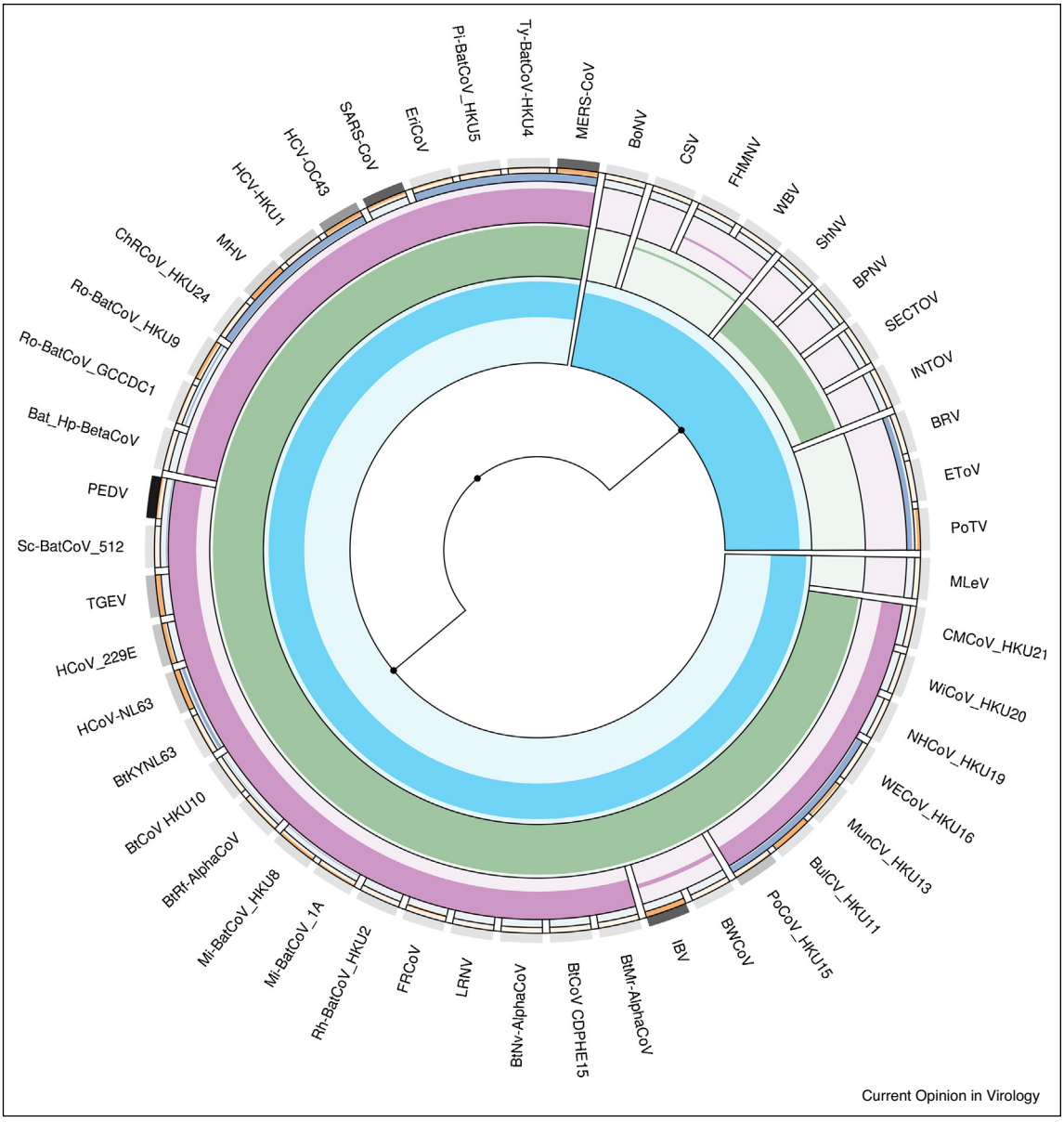
This concern has been addressed in the DEmARC data-driven and phylogeny-aware framework that was developed for the in-depth analysis of large monophyletic groups using ML pairwise distances separating conserved proteins [59]. DEmARC seeks minimizing the cost associated with pairwise distances violating a rank threshold, thereby enabling inference of associated limits on genetic divergence, and includes a check for the monophyly of clusters. The deduced distance thresholds demarcate also taxa that include a single virus, which may be numerous in large virus families due to an extremely unbalanced virus sampling. In principle, DEmARC has the potential to delineate all ranks in a classification of a large monophyletic group. A different and similarly sophisticated approach to defining rank thresholds using pair-wise distances was realized in the VICTOR package [60]. It involves calculation of complex BLAST-based intergenomic distances with the GBDP tool [61] and feedback from the OPTSIL software [62] that maximizes agreement between phylogeny-based clustering of analyzed viruses and taxa at a chosen rank of a reference classification, from species to family. VICTOR has been developed to assist with classification of prokaryotic and archaeal viruses and does not require monophyly of the trunk of the dataset tree, since its taxa may be based on different genomic regions. If distance thresholds are known, the researcher may infer virus clustering using BLAST intergenomic scores calculated by VIRIDIC [63]. For coronaviruses, BLAST scores for partial RdRp gene sequences may be sufficient for virus assignment to known subgenera by MyCOV [64]. All tools listed above can assist experts in developing virus taxonomies within the conventional framework.

### Natural origins of taxa at different ranks

Most computational tools of virus taxonomy that delineate taxa at the species rank are based on analyses of pairwise distances (Table 1 and Figure 1). At least two tools, CASTOR [65] and SDT [58•], were also used for subspecies delineation, which is the domain of specialized software, for example, PhyloPlace [66] and PhyCliP [67], as well as applications of other tools, for example, APCluster [68], which is based on the affinity propagation algorithm popular in image recognition [69]. We included these software packages in our review for two reasons: (a) the used algorithms may be informative for future development of virus taxonomy techniques, even if they may not be applicable directly; (b) it would be interesting to see whether their use could be extended to taxa at the species rank. Indeed, subspecies tools, and especially PhyCliP [67], are particularly robust in accommodating peculiarities of the fast evolution of viruses and uneven virus sampling, which is limited to the most recent decades. These specifics impede accurate estimation of the substitution accumulation that may deteriorate extremely fast with increased sequence divergence.

Verification of species assignment may support recognition of virus species as a population-based entity, similar to contemporary circulating lineages [20,70•].This notion is a bedrock in cellular biology but remains peripheral in virology, leading to persistent underappreciation of virus species outside virus taxonomy [22•]. For instance, naming a human pathogen based on its membership in a known species rather than after the associated disease was unheard in virology till SARS-CoV-2 [52•], whose species-based name caused considerable bewildering. To bridge this divide, vastly expanded intra-species sampling is required. It will improve certainty regarding demarcation of many species by bioinformatics, as we argued for filoviruses [71], with potentially broad implications for census of virus diversity [22•] and beyond. The large-scale genomic characterization may also address a principal concern about delimitation of species and other taxa: does it reflect undersampling from a much wider and dense continuum of the natural virus diversity [7], or capture genuine fractality that emerged due to action of various biological and environmental forces affecting virus diversity directly or through hosts [70•,72]. Under the later hypothesis, stability of the delineated taxa in the face of novel virus discovery may be expected. It's been the case with the DEmARC-defined taxa within the family *Coronaviridae* over more than 10 years (https://ictv.global/taxonomy/).

**Figure 2**




Circular representation of the five-ranks taxonomy of the families *Coronaviridae* and *Tobaniviridae*, along with the density of virus sampling per species.

Intervirus genetic divergence increases linearly from the perimeter toward the center of the circle. The delimited taxa are depicted in rectangle-like shapes that are colored in rank-specific manner: species (orange), subgenus (blue), genus (purple), subfamily (green), and family (marine). Each color may be presented in two shadings, soft or bright, that highlight the limit on intragroup genetic divergence according to a distance threshold (soft) and the maximum observed intra-rank genetic divergence of a taxon (bright). On top of the species bars, the relative density of virus sampling per species as of 2019 is shown as gray shadings from low (light) to high (dark) sampling, which is in the range of 1 (least sampled species) to 365 (most sampled species, PEDV). Each species is labelled with the name acronym of a representative virus. Taxa conform to rank distance thresholds common for the two families (equal, rank-specific heights of taxon shapes). Adapted from Ref. [3].

## Visualization of virus taxonomy

Besides specific computational tools and methods, results visualization is integral to field identity. Virus taxonomy may be presented in a tabular format or using depictions that were borrowed from other fields. Presenting a taxonomic classification next to tips of a phylogenetic tree conveys position, scope, and relations between different taxa; it is the default choice in the 10th ICTV Report and an integral component of any taxonomic proposal. Moreover, practitioners increasingly favor heat maps, a generic matrix depiction of clustering that is readily generated by several taxonomy programs, for example, SDT.

Presenting density distributions of pair-wise distances is also common for visualization of the classification rank structure in distance-based analyses. Each of these depictions are informative, although not uniquely associated with the taxonomy framework to assert its identity. To address this challenge, an original circular plot of virus classification was devised in the DEmARC package [70•]. It depicts sampling density of each species and relationship between taxa at all ranks in distance space along with virus phylogeny for taxa at the basal rank (Figure 2) and can be used to compare classifications. A somewhat similar circular depiction was also proposed by the VICTOR authors (see Figure 2 in Ref. [60]) and used in phylogeny-based classification by [50•].

## Concluding remarks

A growing number of computational tools from comparative genomics have been developed recently to meet the formidable challenges that virus taxonomy is facing in an era of viral metagenomics and high-throughput virus discovery. These challenges are defined by a vast number of unclassified viruses and the profound specifics of virus evolution (see Introduction). They are further amplified by the enormity of the ICTV ambition to classify *all* viruses within a *common* framework. It is in contrast to the practice in taxonomies of host organisms, for example, plants, animals, bacteria and so on, that affect the development of taxonomies of the respective virus groups. To overcome these divides, the development and benchmarking of innovative software must be accompanied with training how to use these tools practically. Fostering bioinformatics-led research on virus classification should include the testing of alternative classifications produced with different tools [50•], as a way to assure the quality of any new taxonomy, resolve disputes between rivalling taxonomy proposals and improve the connectivity to other fields, including metagenomics and experimental testing of taxonomy-inspired hypotheses. These future advancements may promote appreciation of taxonomy by researchers and address the ultimate question about genuine foundations of species and other taxa in virology.

## Conflict of interest statement

Nothing declared.

## Acknowledgements

We thank Stuart Siddell for fruitful collaboration on 'Taxonomy of Viruses', which was instrumental for this project, and an anonymous reviewer for suggestions. AEG work cited in this review was supported by grants from Leiden University Medical Center, Leiden University Fund, and the EU FP and Horizon2000 Programs. CL is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2155 – project number 390874280. AEG and CL are members of the European Virus Bioinformatics Center and served for the ICTV in different capacities.

## References and recommended reading

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

- • of special interest

1. Brister JR, Ako-adjei D, Bao YM, Blinkova O: **NCBI viral genomes resource**. *Nucleic Acids Res* 2015, **43**:D571-D577.

2. *Virus Bioinformatics*. Frishman D, Marz M.Boca Raton: CRC Press; 2021.

3. Gorbalenya AE, Lauber C, Siddell S: **Taxonomy of viruses**. *Reference Module in Biomedical Sciences*. Elsevier; 2019.

4. Kuhn JH: **Virus taxonomy**. *Encyclopedia of Virology*. 2021:28-37.

5. • Adams MJ, Lefkowitz EJ, King AM, 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

This paper and that by Simmonds *et al.* below bring together members of the ICTV-EC and leading experts in (comparative) virus genomics to endorse the genome sequence as a sufficient basis for recognition and classification of new viruses within the virus taxonomy framework.

6. Edwards RA, Rohwer F: **Viral metagenomics**. *Nat Rev Microbiol* 2005, **3**:504-510.

7. Zhang Y-Z, Chen Y-M, Wang W, Qin X-C, Holmes EC: **Expanding the RNA virosphere by unbiased metagenomics**. *Annu Rev Virol* 2019, **6**:119-139.

8. Wolf YI, Silas S, Wang YJ, Wu S, Bocek M, Kazlauskas D, Krupovic M, Fire A, Dolja VV, Koonin EV: **Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome**. *Nat Microbiol* 2020, **5**:1262-1270.

9. • 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

This paper and that by Adams *et al.* above bring together members of the ICTV-EC and leading experts in (comparative) virus genomics to endorse the genome sequence as a sufficient basis for recognition and classification of new viruses within the virus taxonomy framework.

10. • Gorbalenya AE, Krupovic M, Mushegian A, Kropinski AM, Siddell SG, Varsani A, Adams MJ, Davison AJ, Dutilh BE, Harrach B *et al.*: **The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks**. *Nat Microbiol* 2020, **5**:668-674

Prompted by the advent of metagenomics and success of comparative genomics, the ICTV-EC introduces a Linnaean-like rank structure into the virus taxonomy that reinforces a link between virology and cellular biology.

11. Siddell SG, Walker PJ, Lefkowitz EJ, Mushegian AR, Adams MJ, Dutilh BE, Gorbalenya AE, Harrach B, Harrison RL, Junglen S *et al.*: **Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018)**. *Arch Virol* 2019, **164**:943-946.

12. 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.

13. King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (Eds): *Virus Taxonomy, Classification and Nomenclature of Viruses. Ninth Report of the International Committee on Taxonomy of Viruses*. Amsterdam: Elsevier, Academic Press; 2012 https://talk.ictvonline.org/ictv-reports/ictv_9th_report/.

14. Pace NR, Sapp J, Goldenfeld N: **Phylogeny and beyond: scientific, historical, and conceptual significance of the first tree of life**. *Proc Natl Acad Sci U S A* 2012, **109**:1011-1018.

15. • 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-e00019

Pioneering assignment of known viruses to taxa at the newly established ranks basal to the rank order, collectively called megataxonomy; it is based on analysis of remote sequence relationships of most conserved proteins.

16. Lauber C, Seitz S, Mattei S, Suh A, Beck J, Herstein J, Borold J, Salzburger W, Kaderali L, Briggs JAG *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**:387-399.e6.

17. Theze J, Bezier A, Periquet G, Drezen JM, Herniou EA: **Paleozoic origin of insect large dsDNA viruses**. *Proc Natl Acad Sci U S A* 2011, **108**:15931-15935.

18. Hayman DTS, Knox MA: **Estimating the age of the subfamily *Orthocoronavirinae* using host divergence times as calibration ages at two internal nodes**. *Virology* 2021, **563**:20-27.

19. Ghafari M, Simmonds P, Pybus OG, Katzourakis A: **A mechanistic evolutionary model explains the timedependent pattern of substitution rates in viruses**. *Curr Biol* 2021, **31**:1-8.

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

21. Van Regenmortel MHV: **The species problem in virology**. *Adv Virus Res* 2018, **100**:1-18.

22. • Gorbalenya AE, Siddell SG: **Recognizing species as a new focus of virus research**. *PLoS Pathog* 2021, **17**
A case is made for virus species to represent virus taxonomy in communications, as a first step towards recognition of this fundamental unit of biological evolution in virus research and its applications.

23. De Queiroz K: **Species concepts and species delimitation**. *Syst Biol* 2007, **56**:879-886.

24. Mallo D, Posada D: **Multilocus inference of species trees and DNA barcoding**. *Philos Trans R Soc Lond Ser B Biol Sci* 2016, **371**.

25. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB: **Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)**. *Nucleic Acids Res* 2018, **46**:D708-D717.

26. Simmonds P, Aiewsakun P: **Virus classification - where do you draw the line?** *Arch Virol* 2018, **163**:2037-2046.

27. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, Karsch-Mizrachi I: **GenBank**. *Nucleic Acids Res* 2021, **49**:D92-D96.

28. Yu C, Hernandez T, Zheng H, Yau S-C, Huang H-H, He RL, Yang J, Yau SST: **Real time classification of viruses in 12 dimensions**. *PLoS One* 2013, **8**:e64328.

29. Li Y, Tian K, Yin C, He RL, Yau SST: **Virus classification in 60-dimensional protein space**. *Mol Phylogenet Evol* 2016, **99**:53-62.

30. Goldbach RW: **Molecular evolution of plant RNA viruses**. *Annu Rev Phytopathol* 1986, **24**:289-310.

31. Strauss JH, Strauss EG: **Evolution of RNA viruses**. *Annu Rev Microbiol* 1988, **42**:657-683.

32. Gorbalenya AE, Koonin EV: **Comparative analysis of the amino acid sequences of the key enzymes of the replication and expression of positive-strand RNA viruses. Validity of the approach and functional and evolutionary implications**. *Sov Sci Rev D Physicochem Biol* 1993, **11**:1-84.

33. Koonin EV, Dolja VV: **Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences [published erratum appears in Crit Rev Biochem Mol Biol 1993;28(6):546]**. *Crit Rev in Biochem Mol Biol* 1993, **28**:375-430.

34. Le Gall O, Christian P, Fauquet CM, King AMQ, Knowles NJ, Nakashima N, Stanway G, Gorbalenya AE: **Picornavirales, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T = 3 virion architecture**. *Arch Virol* 2008, **153**:715-727.

35. Cavanagh D: ***Nidovirales*: a new order comprising *Coronaviridae* and *Arteriviridae***. *Arch Virol* 1997, **142**:629-633.

36. de Vries AAF, Horzinek MC, Rottier PJM, de Groot RJ: **The genome organization of the *Nidovirales*: similarities and differences between Arteri-, Toro-, and Coronaviruses**. *Semin Virol* 1997, **8**:33-47.

37. 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.

38. • Bin Jang H, 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
Description of the most advanced network-based vConTACT v.2.0 software and its application that demonstrated scalability and processivity for automatic assignment of thousands of DNA phages to taxa at genus to family ranks.

39. 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.

40. Aiewsakun P, Adriaenssens EM, Lavigne R, Kropinski AM, Simmonds P: **Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy**. *J Gen Virol* 2018, **99**:1331-1343.

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

42. Hennig W: **Phylogenetic systematics**. *Annu Rev Entomol* 1965, **10**:97-116.

43. Vences M, Guayasamin JM, Miralles A, De La Riva I: **To name or not to name: criteria to promote economy of change in Linnaean classification schemes**. *Zootaxa* 2013, **3636**:201-244.

44. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0**. *Syst Biol* 2010, **59**:307-321.

45. Price MN, Dehal PS, Arkin AP: **FastTree 2-approximately maximum-likelihood trees for large alignments**. *PLoS One* 2010, **5**.

46. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014, **30**:1312-1313.

47. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ: **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies**. *Mol Biol Evol* 2015, **32**:268-274.

48. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees**. *Bioinformatics* 2001, **17**:754-755.

49. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A: **Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10**. *Virus Evol* 2018, **4**.

50. • 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
The first example of cross-examination of virus bioinformatics tools for classification of a very large and diverse group of phages at the genus to family ranks using a phylogenetic framework common outside virology.

51. Lauber C, Ziebuhr J, Junglen S, Drosten C, Zirkel F, Nga PT, Morita K, Snijder EJ, Gorbalenya AE: **Mesoniviridae: a proposed new family in the order *Nidovirales* formed by a single species of mosquito-borne viruses**. *Arch Virol* 2012, **157**:1623-1628.

52. • Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Lauber C, Leontovich AM, Neuman BW *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
By breaking with a century-old tradition, the *Coronaviridae* Study Group names a human virus pathogen after its species, which was demarcated by comparative genomic within the DEmARC framework.

53. Modha S, Thanki AS, Cotmore SF, Davison AJ, Hughes J: **ViCTree: an automated framework for taxonomic**

**classification from protein sequences**. *Bioinformatics (Oxford, England)* 2018, **34**:2195-2200.

54. Kapli P, Lutteropp S, Zhang J, Kobert K, Pavlidis P, Stamatakis A, Flouri T: **Multi-rate poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo**. *Bioinformatics (Oxford, England)* 2017, **33**:1630-1638.

55. Felsenstein J: *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, Inc.; 2004.

56. Shukla DD, Ward CW: **Amino acid sequence homology of coat proteins as a basis for identification and classification of the potyvirus group**. *J Gen Virol* 1988, **69**:2703-2710.

57. • 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
Very popular web-based tool for fast approximate classification of virus genome sequences using predefined rank thresholds that are imposed on the density distribution of pair-wise distances.

58. • Muhire BM, Varsani A, Martin DP: **SDT: a virus classification tool based on pairwise sequence alignment and identity calculation**. *PLoS One* 2014, **9**:e108277
Very popular virus taxonomy software for generation of a pair-wise distance matrix to assist with hierarchical classification of the analyzed viruses.

59. 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.

60. Meier-Kolthoff JP, Goker M: **VICTOR: genome-based phylogeny and classification of prokaryotic viruses**. *Bioinformatics* 2017, **33**:3396-3404.

61. Auch AF, Klenk HP, Goker M: **Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs**. *Stand Genomic Sci* 2010, **2**:142-148.

62. Goker M, Garcia-Blazquez G, Voglmayr H, Telleria MT, Martin MP: **Molecular taxonomy of phytopathogenic fungi: a case study in peronospora**. *PLoS One* 2009, **4**.

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

64. Wilkinson DA, Joffrin L, Lebarbenchon C, Mavingui P: **Analysis of partial sequences of the RNA-dependent RNA polymerase gene as a tool for genus and subgenus classification of coronaviruses**. *J Gen Virol* 2020, **101**:1261-1269.

65. Remita MA, Halioui A, Diouara AM, Daigle B, Kiani G, Diallo AB: **A machine learning approach for viral genome classification**. *BMC Bioinformatics* 2017, **18**.

66. Hraber P, Kuiken C, Waugh M, Geer S, Bruno WJ, Leitner T: **Classification of hepatitis C virus and human immunodeficiency virus-1 sequences with the branching index**. *J Gen Virol* 2008, **89**:2098-2107.

67. Han AX, Parker E, Scholer F, Maurer-Stroh S, Russell CA: **Phylogenetic Clustering by Linear Integer Programming (PhyCLIP)**. *Mol Biol Evol* 2019, **36**:1580-1595.

68. Fischer S, Freuling CM, Muller T, Pfaff F, Bodenhofer U, Hoper D, Fischer M, Marston DA, Fooks AR, Mettenleiter TC *et al.*: **Defining objective clusters for rabies virus sequences using affinity propagation clustering**. *PLoS Neg Trop Dis* 2018, **12**.

69. Bodenhofer U, Kothmeier A, Hochreiter S: **APCluster: an R package for affinity propagation clustering**. *Bioinformatics* 2011, **27**:2463-2464.

70. • 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
Conceptualization, evaluation and visualization of a hierarchical classification that is based on clustering analysis of pair-wise evolutionary distances between proteins universally conserved in a virus family and produced by DEmARC [59].

71. Lauber C, Gorbalenya AE: **Genetics-based classification of filoviruses calls for expanded sampling of genomic sequences**. *Viruses Basel* 2012, **4**:1425-1437.

72. Wang LF, Walker PJ, Poon LLM: **Mass extinctions, biodiversity and mitochondrial function: are bats' special' as reservoirs for emerging viruses?** *Curr Opin Virol* 2011, **1**:649-657.