Chapter 3

Bacteriophage Taxonomy: A Continually Evolving Discipline

Dann Turner, Evelien M. Adriaenssens, Susan M. Lehman, Cristina Moraru, and Andrew M. Kropinski

Abstract

While taxonomy is an often underappreciated branch of science, it serves very important roles. Bacteriophage taxonomy has evolved from a discipline based mainly on morphology, characterized by the work of David Bradley and Hans-Wolfgang Ackermann, to the sequence-based approach that is taken today. The Bacterial Viruses Subcommittee of the International Committee on Taxonomy of Viruses (ICTV) takes a holistic approach to classifying prokaryote viruses by measuring overall DNA and protein similarity and phylogeny before making decisions about the taxonomic position of a new virus. The huge number of complete genomes being deposited with the National Center for Biotechnology Information (NCBI) and other public databases has resulted in a reassessment of the taxonomy of many viruses, and the future will see the introduction of new viral families and higher orders.

Key words ICTV, NCBI, Taxonomy, Morphology, DNA sequence homology

1 Why Is Taxonomy Important?

Humans like to put things into boxes and then give those boxes names. This process provides both a context, this thing is like these others, and a language—together these things are called something. Unsurprisingly, the art and science of grouping things is particularly important in biology, as it provides a basis for the identification and inference of relationships. In this context, taxonomy is the process of establishing criteria for the contents of individual boxes and a consistent framework that unites them.

The defining characteristic of virus taxonomy is a group of concepts that can be assembled into a hierarchy. Each layer of this hierarchy must be defined so that a species is defined as one thing and a genus is defined as a higher-order thing that encompasses all of the species beneath it. One can imagine a variety of criteria that could be used to define a taxonomic hierarchy, and the preferred set

Joana Azeredo and Sanna Sillankorva (eds.), *Bacteriophage Therapy: From Lab to Clinical Practice*, Methods in Molecular Biology, vol. 2734, https://doi.org/10.1007/978-1-0716-3523-0_3,

of taxonomic metrics is dependent on the nature of relationships under scrutiny and the availability of data that can be used in evaluations.

Setting criteria for these taxonomic definitions is really where art meets science. Taxonomic hierarchies often attempt to address longer-range evolutionary relationships that stretch back into the far-distant past, yet the process of assignment is restricted to observations gleaned from the extant data. So, even under the best conditions, when there is plentiful data to evaluate, taxonomic constructs often require inferences beyond available data. Making matters worse, in the context of bacteriophage taxonomy, there has been a shortage of data to evaluate, making it difficult to both establish unifying taxonomic criteria and to extrapolate a taxonomic framework from these criteria. The rise in the number of high-throughput sequence-based studies, including metagenomics, has added a wealth of data which is being used to create more robust phylogenies and taxa [1, 2].

2 Brief History of Phage Taxonomy Prior to 2008

The formal taxonomy of the tailed bacteriophages originated in the pioneering phage classification work of David Bradley (Memorial University, Canada), who used electron microscopy and acridine orange staining to classify these viruses into three morphotypes: A (contractile tail), B (long noncontractile tail), and C (short noncontractile tail) [3, 4]. This system was adopted and extended in 1971 by the then International Committee on Nomenclature of Viruses (ICNV), with the names *Myoviridae, Styloviridae*, and *Podoviridae* proposed for the three morphotypes by Hans-Wolfgang Ackermann and Abraham Eisenstark of the Bacterial Virus Subcommittee in 1975 [5, 6]. The names of these families were accepted by the International Committee on Taxonomy of Viruses (ICTV) in 1981 (as *Myoviridae* and *Podoviridae*; https://ictv.global/taxonomy/history) and in 1984 (*Siphoviridae*; https://ictv.global/taxonomy/history). In 1998, Ackermann proposed an order, the *Caudovirales* [7] to encompass all the tailed phages, which was approved by a postal vote that year. The classification system at the family level remained largely unchanged for 37 years, and it provided an invaluable framework for the classification of bacterial viruses in the absence of sequence data.

The advent of the "omics era" coupled with renewed interest in bacterial viruses has had a profound effect on phage classification. A seminal paper on phage evolution was published in 1999 by Roger W. Hendrix and colleagues [8]. In their paper, subtitled "All the World's a Phage," the authors argue cogently that "all dsDNA phage genomes are mosaics with access, by horizontal exchange, to a large common genetic pool but in which access to the gene

pool is not uniform for all phage." This led to a period in which little advancement was made in official phage taxonomy because it was considered that rampant recombination would blur taxa boundaries.

The next major advance in phage grouping occurred with the publication of the highly controversial "The Phage Proteomic Tree: A Genome-Based Taxonomy for Phage" in which Forest Rohwer and Rob Edwards employed BLASTp to compare the proteomes of 105 fully sequenced phage genomes [9]. It was controversial because of some of the illustrated relationships: The "P2 Myophage" cluster contained coliphages P2 and Mu; the "PZA Podophage" group included a member of the *Tectiviridae* (coliphage PRD1) and *Bacillus* podoviruses PZA and GA-1; and, lastly the "λ-like Siphophage" cluster harbored *Escherichia coli* phage λ, HK97 and 933 W, *Pseudomonas* phage D3, and *Salmonella* phage P22. There is no doubt that the latter phages share important features with λ [10], but the merging of members of the *Siphoviridae* and *Podoviridae* created an intellectual stir.

2.1 Extension of Proteomics to Phage Taxonomy from 2008

From 2002 to 2008, the number of fully sequenced members of the *Caudovirales* in GenBank had increased from 19 to 276. However, the number of ICTV-classified species remained at 36. In 2008, when Rob Lavigne (KU Leuven, Belgium) assumed the Chair of ICTV's Bacterial and Archaeal Viruses Subcommittee, it was time to take a new look at the classification of phages. Using two protein analysis tools, CoreExtractor and CoreGenes, Lavigne and co-workers realized that the T7-like phages actually fell into three distinct clades, which were termed "T7-like virus," "SP6-like virus," and "φKMV-like virus" each containing multiple species [11]. Since these three groups shared a similar genomic organization and the presence of a large, single subunit RNA polymerase, they were grouped into the first phage subfamily, *Autographivirinae*. What linked the species within a clade was that the members shared 40% of their proteins in common, as shown using CoreGenes [12–14]. This threshold was defined based on a comparison of T7 and *Pseudomonas* phage gh-1 [15], during which detailed molecular analysis revealed that this phage was closely related to *Escherichia coli* phage T7. This approach was subsequently used with the *Myoviridae* [16] and *Siphoviridae* [17]. The problem with this total proteome approach is that the genomes being compared have to be fully and correctly annotated, which is not always the case. In addition, CoreGenes is relatively slow and tedious to apply to multiple genomes. The latest version, CoreGenes5.0, allows the determination of the common proteins encoded by a maximum of 21 phages [14].

In 2008, another seminal paper was published recognizing the reticulate nature of phage genomic relationships using a network-based representation of phage populations [18, 19]. The findings

and progress made in this paper were largely ignored for a decade until other researchers used similar approaches to analyze a larger dataset of the dsDNA virosphere, including archaeal viruses [20, 21]. Not until the development of vConTACT and its successor vConTACT2, which use similar gene-sharing networks as first used by Gipsi Lima-Mendez and colleagues, did network approaches to represent phage diversity become really popular [22, 23]. The downside of these approaches is that they need to be converted into a hierarchical classification before they can be used in phage taxonomy.

2.2 DNA Sequence Comparisons Enter the Picture

DNA-DNA sequence relatedness has always been the gold standard for the classification of bacterial strains [24, 25], and while DNA-DNA hybridizations have been used to study phage relationships [26–28], in silico analyses of the sequence relationships between biological entities did not influence phage taxonomy until fairly recently. The most commonly used genome comparison tools were progressiveMauve [29], EMBOSS Stretcher [30, 31], FastANI [32], and dot matrix analysis programs [33–35]. The problem with EMBOSS Stretcher and FastANI is that they are inaccurate below approximately 50% sequence identity, they require collinear genomes, and can only handle pairs of viruses. Other lesser used graphical comparison tools that provide metrics of sequence similarity include BRIG (BLAST Ring Image Generator; [36]), Easyfig [37], Circos [38], CGView [39], and CGView Comparison Tools [40]. Dotplots have been used extensively by scientists associated with the Actinobacteriophage Database projects [41–43], and an example Gepard plot ("**GE**nome **PA**ir – **R**apid **D**otter"; [35]; http://cube.univie.ac.at/gepard) comparing two *Leuconostoc* phage genomes is shown in Fig. 1.

Though progressiveMauve, EasyFig, BRIG, etc., and all of the dot matrix analysis tools produce good figures for manuscripts, they do not express relationships between genomes in quantitative terms such as "percent identity." This is one of the advantages of EMBOSS Stretcher, PASC [PAirwise Sequence Comparison; [44]; (https://www.ncbi.nlm.nih.gov/sutils/pasc/viridty.cgi)], SDT [Sequence Demarcation Tool; [45]; (http://web.cbio.uct.ac.za/~brejnev/)], JSpecies ([46]; https://imedea.uib-csic.es/jspecies/), ANI (Average Nucleotide Identity; [47]; http://enve-omics.ce.gatech.edu/ani/), GGDC (Genome-To-Genome Distance Calculator; [48]), and VICTOR ([49]; https://ggdc.dsmz.de/). However, with the possible exception of ANI and VICTOR, most of these tools have not been used for phage research. To be more precise, VICTOR calculates pairwise identity scores and uses them to calculate the tree and the different taxon levels, but does not output the "percentage identity."

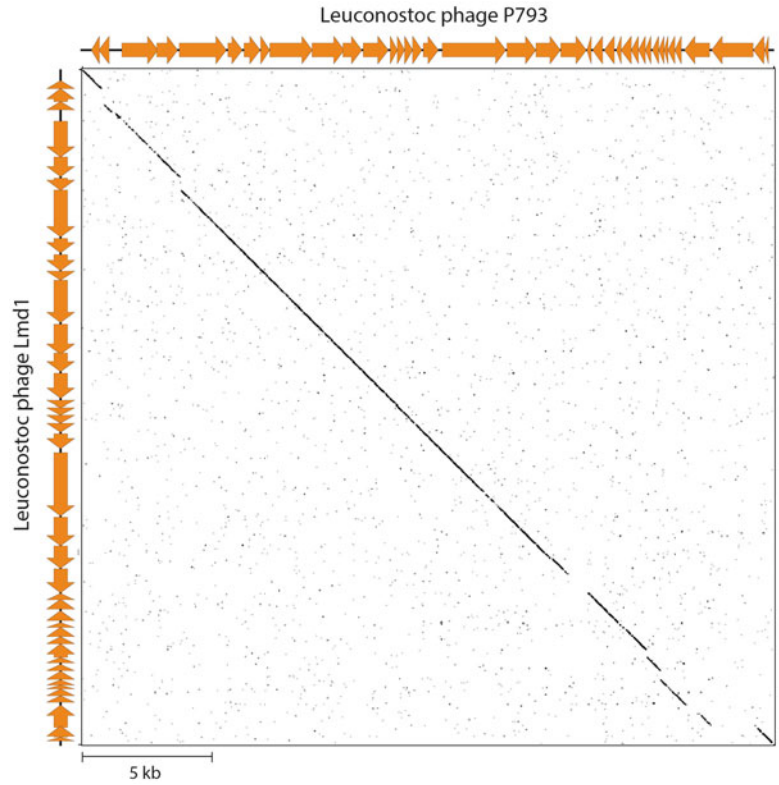

Fig. 1 Gepard dotplot comparing the similarity between the genomes of *Leuconostoc* siphoviruses Lmd1 [93] versus P793 [94] modified to include genome maps generated using EasyFig. Protein-coding genes are depicted as arrows with the orientation indicating the location on the forward or reverse strand. The dark diagonal lines represent syntenic regions

Two groups have made extensive use of DNA sequence homology to group bacteriophages. The first is the Actinobacteriophage Database (https://phagesdb.org/), which currently includes 20,918 phages [50]. The original clustering method [51] has now been modified to include a measure of the Gene Content Similarity (GCS; [52]). Second, in 2014 Grose and Casjens [53, 54] used BLASTn, BLASTp, and quantitative DotPlot data to group 337 sequenced Enterobacterial phages into 56 clusters, many of which now correspond to ICTV-ratified genera.

3 How ICTV Currently Groups Phages into Taxa

It is imperative before continuing this discussion to distinguish between phage isolates and taxa. To quote from the ICTV website, "Viruses are real physical entities produced by biological evolution and genetics, whereas virus species and higher taxa are abstract

concepts produced by rational thought and logic. The virus/species relationship thus represents the front line of the interface between biology and logic"" [55–57]. In a recent opinion piece, this is further clarified as "virus species are human-made taxonomic categories to which viruses are assigned when they satisfy a particular set of properties, known as "species demarcation criteria"" [55, 58]. In practical terms, phage isolates are physical entities that can form plaques on an appropriate host while taxa cannot. Once a new phage is isolated, it can be added to an existing taxon, or in cases where the new isolate is sufficiently distinct from extant isolates, a new taxon can be proposed with the new isolate the exemplar of the new species. The tool for proposing the creation of a new viral order, family, subfamily, genus, and species, as well as for modifying existing taxa, is known as a Taxonomy Proposal (abbreviated as TaxoProp). The template for this can be downloaded from the appropriate ICTV website (https://ictv.global/taxonomy/templates). Anyone can fill in and submit a TaxoProp, but it is generally advisable to work with an appropriate member of the Bacterial Viruses Subcommittee (BVS; https://ictv.global/sc/bacterial), who can offer advice and proofread the intended submission.

The process of placing a new phage isolate within a taxon begins with identifying all related isolates. This can be done by comparing the new isolate sequence to the "nucleotide collection (nt/nr)" database using the BLASTn algorithm [59]. Searches should be restricted to Organism "Viruses (taxid:10239)" or "Caudoviricetes (taxid:2731619)," if the morphology is known. The increase in SARS-CoV-2 sequences means BLASTn often returns a CPU overload message now. To prevent identifying prophages that are part of bacterial genomes, the search is limited to Viruses, unless prophages are what you are interested in. It is also helpful to search against the reference genomic sequences (refseq_genomic) database, as new viral and phage reference genomes are created for each species exemplar [60–62]. Alternatively, one can search against a specific phage species, genus, subfamily, or family. A crude estimate of the overall nucleotide sequence similarity between pairs of genomes can be obtained from the BLASTn search results by multiplying the "Query Cover" by "Per. Ident."

Turner et al. [63] have created "A Roadmap for Genome-Based Phage Taxonomy" in which the demarcation criteria necessary to create the following taxa are enunciated. In the following taxonomic sections, we have quoted liberally from this manuscript.

Species: "The biological species concept [. . .] defines species as interbreeding individuals that remain reproductively isolated from other such groups" (cited from [64]). In the absence of sexual reproduction, gene flow, for example through homologous recombination, has been proposed to result in "interbreeding" [65]. Therefore, viral species can be thought of as a group of strains

with high rates of gene exchange. In nature, this would result in the existence of sequence-discrete populations. A large-scale metagenomic study in the surface oceans found that dsDNA phage populations form discrete genotypic clusters at ~95% average nucleotide identity (ANI) cut-off [64].

This is in line with the current ICTV species threshold. "Two phages are assigned to the same species if their genomes are more than 95% identical over their genome length for isolates." These values can be calculated by a number of tools, such as BLASTn, or using the intergenomic distance calculator VIRIDIC [66]. VIRIDIC calculates nucleotide-based intergenomic identities that are normalized to the viral genome lengths. Therefore, VIRIDIC values are less prone to overestimation of the intergenomic identity as compared with ANI methods, which normalize identity to alignment length. It is worth noting, however, that large direct terminal repeats can still distort VIRIDIC results unless the user ensures that all of the input genomes contain only one copy of the terminal repeat.

In Fig. 2, VIRIDIC has been applied in a quantitative genome comparison of a group of *Leuconostoc* siphoviruses. ICTV does not classify strains, but BVS keeps a list of these and transmits this information to NCBI.

Genus: Turner et al. state: "In search for criteria that create cohesive and distinct genera that are reproducible and monophyletic, the Bacterial Viruses Subcommittee has established 70% nucleotide identity of the genome length as the cut-off for genera." In the case of temperate phages, some leeway to this value is permitted. Genus-level groupings should always be monophyletic in the signature genes, as tested by phylogenetic analysis. In Fig. 2, phage P793 and its homologs are classified into the genus *Limdunavirus*, while phage CHB is a member of the *Unaquatrovirus*. Based upon the above statement concerning strains, phages Diderot and phiLN03 (exemplar) are both strains within the species *Limdunavirus LN03*; phages phiLN6B and Lmd1 (exemplar) within *Limdunavirus Lmd1*; phiLN34 (exemplar), phiLNTR2, P974, CHA, and CHB are strains within *Unaquatrovirus LN34*; and, lastly, Ln-8 and phiLN25 (exemplar) are strains within the species *Unaquatrovirus LN25*.

Subfamily: Subfamilies are to be created when two or more genera are related below the family level. In practical terms, this usually means that they share a low degree of DNA sequence similarity (usually about 40–50%) and that the genera form a clade in a marker tree phylogeny. In the above example, we considered that there was sufficient evidence for a taxonomic relationship between the *Limdunavirus* and the *Unaquatrovirus* to create a subfamily which was named *Mccleskeyvirinae*.

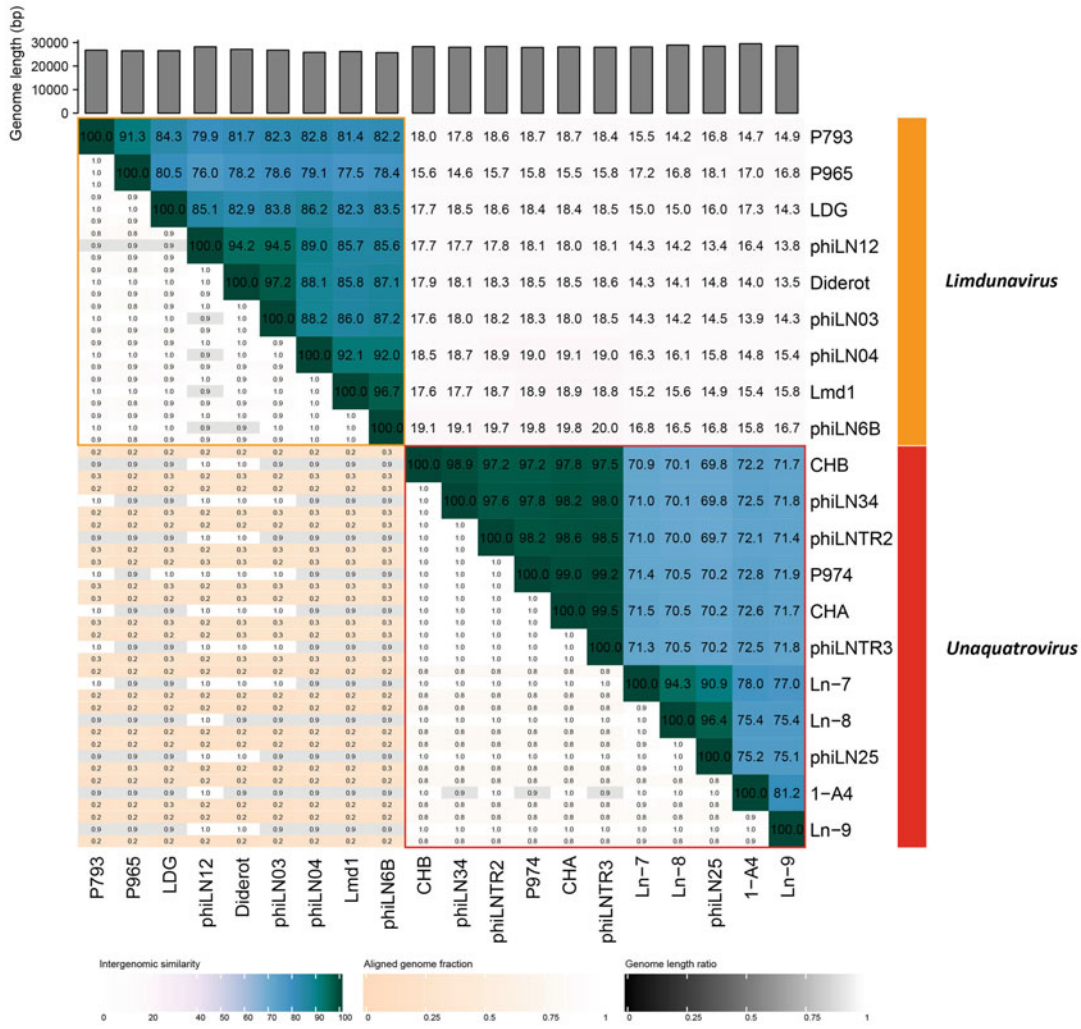

Fig. 2 VIRIDIC analysis of 20 *Leuconostoc* siphoviruses in which two distinct clades (*red square* for *Unaquatrovirus* genus and *yellow square* for *Limdunavirus* genus) are recognizable. The upper-right half shows the nucleic acid intergenomic identities for each genome pair, color coded in a range from white (0% identity) to dark green (100% identity). The lower-left half shows three different indicator values. For each genome pair, these values are given at the intersection of the corresponding row with the corresponding column, as follows (from the top to bottom of the intersection square): (i) aligned fraction genome 1, representing the proportion of the genome found on the row that has been aligned to its pair (the genome on the column); (ii) genome length ratio between the two viruses in a pair; and (iii) aligned fraction genome 2, representing the proportion of the genome found on the column that has been aligned to its pair (the genome on the row). All three values are color coded as well (see legend). Lighter, closer to white colors, indicate well-aligned genome fractions (close to 1) and/or good genome length ratios (close to 1). For example, the pair phiLNTR2 (on the row) and CHB (on the column) has all three indicators equal to 1, meaning that their genome lengths are very similar and that they align along their full genome. On the contrary, the pair phiLNTR2 (row) and phiLN6b (column) have two aligned genome fractions of 0.3 and their genome length ratio of 0.9. This means that even though the lengths of their genomes are similar, they have regions of homology (and therefore align) only on 30% of their respective genome lengths. Their intergenomic identity is also low—19.7

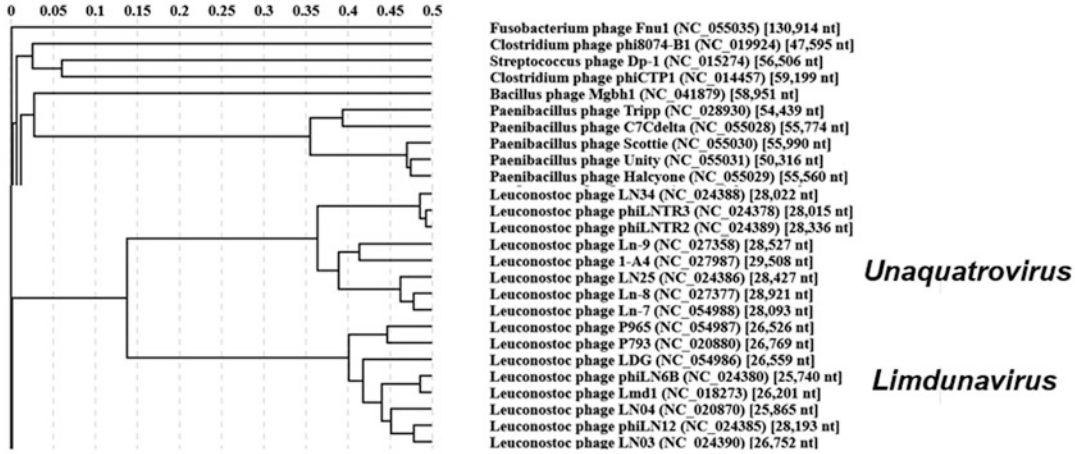

Fig. 3 ViPTree analysis of *Leuconostoc* siphoviruses. ViPTree analysis (https:/www.genome.jp/viptree/) is based upon Rohwer and Edwards's [9] Phage Proteomic Tree and utilizes tBLASTx comparisons between pairs of phage genomes

Family: "The family is represented by a cohesive and monophyletic group in the main predicted proteome-based clustering tools," which can include ViPTree [67], GRAViTy dendrogram [68, 69], vConTACT2 network [22, 23], and VirClust [70, 71]. To add more functional information to these proteome-based clustering tools, we can recommend the PHROGs database (Prokaryotic Virus Remote Homologous Groups; [72]) and Phamerator [73, 74]. PHROGs represents a comprehensive database of viral protein superclusters, which allows HMM comparisons and, thus, enables the detection of more distant homologues and enhances viral genome annotations and the discovery of viral hallmark genes. Members of a viral family share a significant number of orthologous genes (the number will depend on the genome sizes and the number of coding sequences of members of the family). In the case of these *Leuconostoc* phages (Fig. 3), there is insufficient data to support creating a new family.

Since the ViPTree analysis is based upon tBLASTx and not BLASTp, CoreGenes3.5 [75] or CoreGenes5.0 (https://coregenes.ngrok.io/) are frequently employed to compare proteomes. Other command line tools such as Roary [76], PIRATE [77], and PanOCT [78] are also available to facilitate pan-genome analysis. It is important to note that ViPTree does not consider gene order and should be interpreted with caution (ideally by supplementing with another tool), particularly if there is a biologically surprising result. For example, when trying to resolve the classification of lambda-like phages above the subfamily level, we have seen clades of different morphotypes group more closely together than related clades of the same morphotype. When the

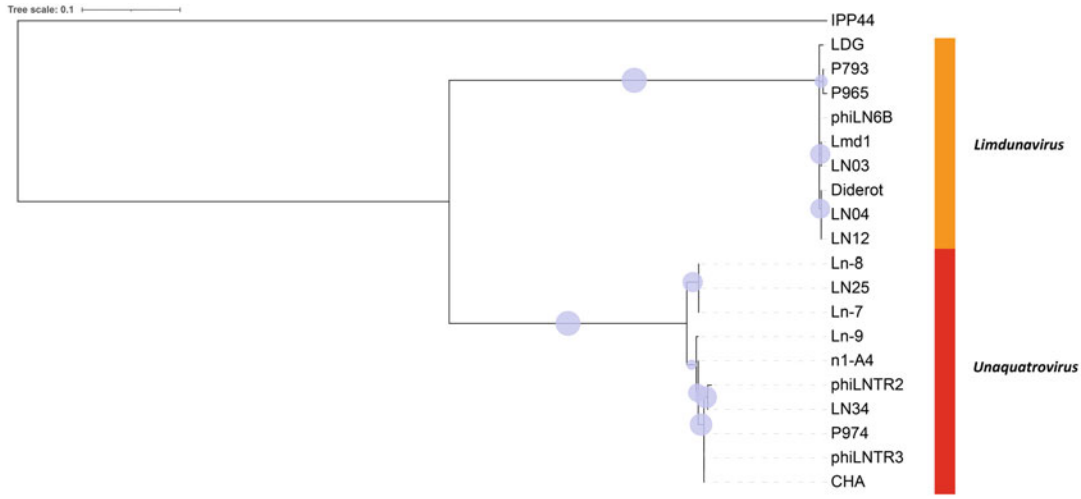

Fig. 4 Large subunit terminase (TerL) phylogenetic tree of the *Leuconostoc* phages with the homologous protein from *Streptococcus* phage IPP44 as the outlier. This was generated using "one click" at http://phylogeny.lirmm.fr/ [95], exported in Newick format, and visualized using iTOL v6 [96]; the bootstrap support >50% is shown as filled circles, with the size proportional to the value. *Leuconostoc* phage CHB is not included since it is represented, in GenBank, as two pieces. "The 'One Click mode' targets users that do not wish to deal with program and parameter selection. By default, the pipeline is already set up to run and connect programs recognized for their accuracy and speed (MUSCLE for multiple alignment and PhyML for phylogeny) to reconstruct a robust phylogenetic tree from a set of sequences." It also includes the use of Gblocks to eliminate poorly aligned positions and divergent regions. "The usual bootstrapping procedure is replaced by a new confidence index that is much faster to compute". (see [97] for further details)

same genomes are run through GRAViTy, which does take genome organization into account, the same-morphotype clades all cluster together.

As the last step in collecting data to support new taxa, phylogenetic trees are created. These are usually based upon virus hallmark genes, for example, the large subunit terminase, DNA polymerase, nucleotide metabolism, major capsid, or major tail proteins. For the family level, the recommendation is to generate a phylogenomic tree, that is, a tree that is based on an alignment of all the core genes in the family. They need to include maximum likelihood, bootstrapping, and rooting using an outgroup to enable an accurate assessment of monophyly here. The results (Fig. 4) of our analysis of TerL proteins from *Leuconostoc* siphophages are completely in accord with the total proteome and total genome analyses. In summary, BVS now employs a more holistic approach to classifying phages, relying on genomic identity along with a common proteome and phylogeny. It should be pointed out that members of the *Autographiviridae*, *Chaseviridae* [79], and *Xipdecavirus* [80] all encode a large single-subunit RNA polymerase gene, yet morphologically represent podo-, myo-, and siphoviruses, respectively.

Table 1
Change in the numbers of various ICTV-recognized taxa belonging to the order *Caudovirales* or class *Caudoviricetes*

Year				
	2005	**2011**	**2016**	**2021**
Class	0	0	0	1
Order	1	1	1	4
Family	3	3	3	47
Subfamily	0	5	15	98
Genus	18	37	215	1197
Species	36	135	956	3601

3.1 Progress

As more and more phage sequences were deposited in databases, it became apparent that the order *Caudovirales* and its three families was not monophyletic and that its existence inhibited the development of a true phage taxonomy. The BVS initially dealt with this problem by creating parallel families, including the *Herelleviridae*, which was used as a case study [81, 82]. However, this proved ultimately unsatisfactory. Therefore, in 2020, ICTV expanded the number of taxonomic ranks from five to fifteen [83] and in 2022 removed the order *Caudovirales*, and the families *Myoviridae*, *Siphoviridae,* and *Podoviridae* [84]. Please note that the terms "myovirus," "podovirus," and "siphovirus" will continue to be acceptable in describing morphotypes.

The BVS of ICTV has made huge advances in method development and in the classification of bacteriophages (Table 1). This is probably best illustrated by the classification of the T7-like phages: In 2008, these were classified as three species in the family *Podoviridae*; as of 2021, they had become the *Autographiviridae* family, with three subfamilies, 133 genera, and 373 species. In the latest taxonomic update, 1197 of the 2606 viral genera are members of what was known as the *Caudovirales*. In the latest taxonomic update, the order *Caudovirales* was abolished and replaced by the class Caudoviricetes to encompass all bacterial and archael viruses with a dsDNA genome. Concerted efforts are now required by both the BVS and wider bacterial virus research community to establish genomically coherent taxa at the levels of family and order. For classification at the family level, ideally, at least two genera comprised of multiple species are required to determine the presence of several hallmark gene products conserved across all species to be used as demarcation criteria.

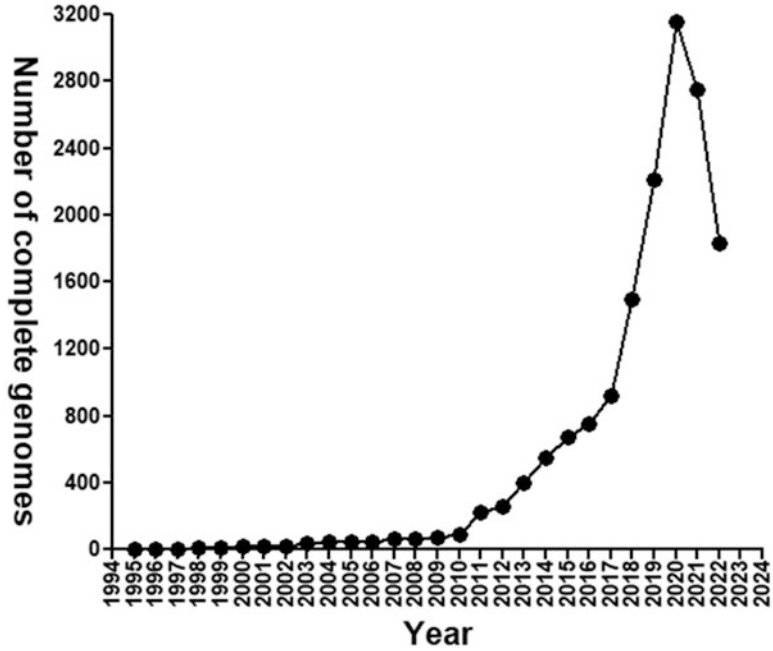

Fig. 5 Annual growth of fully sequenced members of the class *Caudoviricetes* in NCBI. The data is the number of complete *Caudoviricetes* (newer data) or *Caudovirales* (older data) as of September first of each year and was generated using "(((Caudoviricetes) AND complete genome) AND ("2022/08/31" [Publication Date] : "2023/09/01" [Publication Date])) NOT RefSeq"

3.2 Taxonomic Challenges of Modern Sequence Databases

The number of complete phage genomes deposited in the International Nucleotide Sequence Database Collaboration (INSDC) resources annually is staggering (Fig. 5). Since ICTV only reviews TaxoProps once a year and the documentation requires considerable attention to detail, the creation of new taxa lags significantly that of generating complete phage sequences.

Phage sequences in the databases are derived from three sources—direct isolates (phages), bacterial genomes (prophages), and environmental samples (viromes). While there is growing evidence that viromic samples represent complete phages, the same does not extend to prophages, many of which are demonstrably cryptic. In many cases, it is difficult to determine from the online record whether the "phage" was isolated in a laboratory or discovered using bioinformatics approaches. In the case of prophage submissions, the DEFINITION line could be "host genus + prophage + complete genome," or the GenBank sequence record could include the qualifier "/proviral" within the "source" feature key. We would very much like the aforementioned databases to introduce classifications for viral genomes derived from metagenomes and for prophages, to enable the application of filters within BLAST searches. In addition. BLAST searches can reveal "hits" to

metagenome-derived phage-like sequences. A search for "Siphoviridae sp. (human metagenome)" revealed 23,445 sequences that are called "MAG TPA_asm: Siphoviridae sp.," 23,635 deposited under "Siphoviridae sp. (seawater metagenome)," and 24 under "Siphoviridae sp. (animal metagenome)." There are also huge numbers of metagenomically classified *Myoviridae* and *Podoviridae*. Some of those with complete genomes have been classified recently to a new Order, the *Crassvirales*. The presence of those which have not been classified by ICTV can complicate the interpretation of the BLASTn results. If one includes NOT ENV[Division] in the Entrez Query box, all classified environmental (ENV) "hits" will be removed/hidden. Alternatively, we recommend running BLAST searches against the manually curated **IN**frastructure for a **PHA**ge **RE**ference **D**atabase (INPHARED; [85]). Comparisons against the INPHARED database require expertise to create custom BLAST databases for command-line searches.

Incomplete/inaccurate taxonomies and errors in the deposited phage sequences and their annotation contribute to the propagation of errors and can confound classification efforts. In an ideal scenario, submitters should view their sequence record as a living document and take responsibility for updating these records to address errors of annotation or classification as they become apparent.

3.3 Creation of Higher Taxa

It has been proposed that the defining characteristic of the "T4 superfamily" of phages, which infect a wide range of host bacteria in the phyla Proteobacteria and Cyanobacteria, is the presence of approximately 30 conserved proteins [86]. Huge advances have been made in the formalized classification of these viruses by Andrew Millard and his study group, who have created two new families, *Straboviridae* (2 subfamilies, 35 genera) and *Kyanoviridae* (45 genera). These include phages infecting *Aeromonas, Cronobacter, Enterobacter, Escherichia, Erwinia, Klebsiella, Morganella, Pectinobacterium, Proteus, Pseudomonas, Salmonella, Serratia, Shigella, Vibrio*, and *Yersinia*; and *Prochlorococcus* and *Synechococcus*, respectively. *Campylobacter* phages, which by Petrov's definition are T4-like, are to be found in the subfamily *Eucampyvirinae*.

Another age-old taxonomy problem is the relationship between *Escherichia coli* phage λ (*Lambdavirus*) and *Salmonella* phage P22 (*Lederbergvirus*) since these two temperate phages have syntenic genomes and the same type of repressor-anti-repressor regulatory circuitry, but different morphologies. In addition, similarities have been observed with the T1-like phages (*Drexlerviridae*) and N15-like phages (*Ravinvirus*), which share tail morphogenesis genes [87] (http://www.biologyaspoetry.org/PDFs/bgnws018.pdf). While analysis is ongoing, it appears likely that the lower taxonomic groups containing λ, P22, and T1 will remain distinct, with the differences among these groups outweighing their

similarities, but the best approach to structuring these taxa has yet to be determined. At the genus level, in 2020, a large number of unclassified phages that had been described as lambda-like by the historical report, by the sequence submitter, or by NCBI algorithms were collected and placed into 10 genera and 20 species (ICTV TaxoProp 2020.013B). Only one phage presented as a challenging "edge case," having 70–71% nucleotide similarity to all of the phages in each of the two candidate genera. Automated tools slightly favored one genus assignment and manual inspection of the genome alignment slightly favored the other. With no clear "best" result, the phage was ultimately placed within the genus to which it was assigned by automated tools (VIRIDIC, a nucleotide method, and ViPTree, a proteomic method) on the grounds that this is most likely to result in consistent classification by future users who might not inspect every alignment. At the Family, Subfamily, and Genus levels, genome mosaicism has presented some challenges, but does not seem to be the insurmountable challenge that had sometimes been predicted for the historically lambdoid phages.

The last two examples point to the need for the BVS to develop criteria for the identification of higher-order taxons, particularly at this time for families, orders, and classes. Toward this end, one of the authors (CM) is currently developing a web-based tool, VirClust ([72] http://virclust.icbm.de/), aimed to assist researchers with virus classification. VirClust represents a protein-based tool, which calculates protein clusters (based on BLASTp similarity) and protein superclusters (based on HMM similarity) for the input viral genomes. Further, VirClust is using these protein (super)-clusters to (i) cluster hierarchically the input viral genomes, (ii) delineate viral genome clusters based on different intergenomic distance thresholds, (iii) calculate core proteins (signature genes) for the respective viral clusters, and (iv) annotate viral proteins against several databases. To identify higher-order taxons (e.g., Order), it may be necessary to define HMMs of signature genes—see the RdRP of RNA viruses [88].

4 Concluding Statement

Over the last decade, interest in bacteriophages has blossomed. This is a result of three things: (i) the isolation and characterization of phages as alternatives to antibiotics (phage therapy), (ii) the use of these viruses to teach genomics (phage hunting courses; [89–92]), and (iii) the realization that all ecosystems are teaming with phages (metagenomics). Productivity in these areas of phage research has been enhanced by the availability of inexpensive DNA sequencing. Collectively we have seen a massive increase in complete viral genomes in the INSDC databases. At the taxonomic

level, the introduction of an expanded taxonomic field and the removal of the order *Caudovirales* and its three families have and will allow natural evolutionary relationships to be recognized. We have also seen the development of valuable tools for the characterization of metagenomes, along with those which group viruses based on DNA and protein sequence homology.

5 Practical Considerations for Phage Scientists

Before deciding that a newly sequenced phage genome fits within an existing or new genus one should review all the evidence: host; morphotype; lifestyle; genome characteristics (kb, mol% G + C, number of CDS, tRNAs); DNA-DNA relatedness; % protein homologs; and phylogeny. After reviewing all the data, you may decide to lump or split. Do not try and "force" a phage genome into an existing genus, there are plenty of orphan species and genera for which homologs are subsequently isolated. In the case where you experience difficulty, we would recommend that you contact the appropriate member of the Bacterial Viruses Subcommittee (https://ictv.global/sc/bacterial/sc_bacterial).

Acknowledgments

All authors are members of the Bacterial Viruses Subcommittee of the ICTV, with EMA currently its Chair. EMA is supported by the Biotechnology and Biological Sciences Research Council (BBSRC) under the BBSRC Institute Strategic Program Gut Microbes and Health BB/R012490/1 and its constituent projects BBS/E/F/000PR10353 and BBS/E/F/000PR10356. CM's work was supported by the Deutsche Forschungsgemeinschaft within the Transregional Collaborative Research Centre Roseobacter (TRR51).

References

1. Simmonds P, Adams MJ, Benkő M et al (2017) Consensus statement: virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15(3):161–168
2. Callanan J, Stockdale SR, Adriaenssens EM et al (2021) *Leviviricetes*: expanding and restructuring the taxonomy of bacteria-infecting single-stranded RNA viruses. Microbial Genomics 7(11)
3. Bradley DE (1963) The structure of coliphages. J Gen Microbiol 31:435–445
4. Bradley DE (1966) The fluorescent staining of bacteriophage nucleic acids. J Gen Microbiol 44(3):383–391
5. Ackermann HW, Nguyen TM (1983) Sewage coliphages studied by electron microscopy. Appl Environ Microbiol 45(3):1049–1059
6. Ackermann H-W, Nguyen T-M, Delage R (1981) Un nouveau phage d'entérobactéries à tête allongée et queue courte. Ann Inst Pasteur (Paris) 132E:229–234
7. Ackermann H-W (1998) Tailed bacteriophages: the order *Caudovirales*. Adv Virus Res 51:135–201
8. Hendrix R, Smith MC, Burns RN et al (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a

phage. Proc Natl Acad Sci U S A 96(5): 2192–2197

9. Rohwer F, Edwards R (2002) The phage proteomic tree: a genome-based taxonomy for phage. J Bacteriol 184(16):4529–4535. https://doi.org/10.1128/JB.184.16.4529-4535. PMID: 12142423; PMCID: PMC135240
10. Casjens SR, Hendrix RW (2015) Bacteriophage lambda: early pioneer and still relevant. Virology 479–480:310–330
11. Lavigne R, Seto D, Mahadevan P et al (2008) Unifying classical and molecular taxonomic classification: analysis of the *Podoviridae* using BLASTP-based tools. Res Microbiol 159(5): 406–414
12. Kropinski AM, Borodovsky M, Carver TJ et al (2009) *In silico* identification of genes in bacteriophage DNA. Methods Mol Biol 502:57–89
13. Zafar N, Mazumder R, Seto D (2002) CoreGenes: a computational tool for identifying and cataloging "core" genes in a set of small genomes. BMC Bioinform 3(1):12
14. Davis P, Seto D, Mahadevan P (2022) CoreGenes5.0: an updated user-friendly webserver for the determination of core genes from sets of viral and bacterial genomes. Viruses 14(11): 2534
15. Kovalyova IV, Kropinski AM (2003) The complete genomic sequence of lytic bacteriophage gh-1 infecting *pseudomonas putida*-evidence for close relationship to the T7 group. Virology 311(2):305–315
16. Lavigne R, Darius P, Summer EJ et al (2009) Classification of *Myoviridae* bacteriophages using protein sequence similarity. BMC Microbiol 9:224
17. Adriaenssens EM, Edwards R, Nash JH et al (2014) Integration of genomic and proteomic analyses in the classification of the *Siphoviridae* family. Virology 14:10
18. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R et al (2008) Reticulate representation of evolutionary and functional relationships between phage genomes. Mol Biol Evol 25(4):762–777
19. Lima-Mendez G (2012) Reticulate classification of mosaic microbial genomes using NeAT website. Methods Mol Biol 804:81–91
20. Iranzo J, Koonin EV, Prangishvili D et al (2016) Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsidless mobile elements. J Virol 90(24):11043–11055
21. Iranzo J, Krupovic M, Koonin EV (2016) The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. MBio 7(4)
22. Bolduc B, Jang HB, Doulcier G et al (2017) vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect archaea and bacteria. PeerJ 5:e3243
23. Bin Jang H, Bolduc B, Zablocki O et al (2019) Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol 37(6): 632–639
24. Richter M, Rossello-Mora R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 106(45):19126–19131
25. Garrity GM (2016) A new genomics-driven taxonomy of bacteria and archaea: are we there yet? J Clin Microbiol 54(8):1956–1963
26. Ashelford KE, Fry JC, Bailey MJ, Jeffries AR et al (1999) Characterization of six bacteriophages of *Serratia liquefaciens* CP6 isolated from the sugar beet phytosphere. Appl Environ Microbiol 65(5):1959–1965
27. Deveau H, Labrie SJ, Chopin MC et al (2006) Biodiversity and classification of lactococcal phages. Appl Environ Microbiol 72(6): 4338–4346
28. Krylov V, Pleteneva E, Bourkaltseva M et al (2003) *Myoviridae* bacteriophages of *Pseudomonas aeruginosa*: a long and complex evolutionary pathway. Res Microbiol 154(4): 269–275
29. Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5(6):e11147
30. Ceyssens PJ, Glonti T, Kropinski NM et al (2011) Phenotypic and genotypic variations within a single bacteriophage species. Virol J 8:134
31. Rice P, Longden I, Bleasby A et al (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16(6): 276–277
32. Jain C, Rodriguez RL, Phillippy AM et al (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9(1):5114
33. Brodie R, Roper RL, Upton C (2004) JDotter: a Java interface to multiple dotplots generated by dotter. Bioinformatics 20(2):279–281
34. Elnitski L, Riemer C, Schwartz S et al (2003) PipMaker: a World Wide Web server for genomic sequence alignments. Curr Protoc Bioinformatics Chapter 10:Unit 10.2
35. Krumsiek J, Arnold R, Rattei T (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23(8): 1026–1028

36. Alikhan NF, Petty NK, Ben Zakour NL et al (2011) BLAST ring image generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12:402
37. Sullivan MJ, Petty NK, Beatson SA (2011) Easyfig: a genome comparison visualizer. Bioinformatics 27(7):1009–1010
38. Smith DL, Rooks DJ, Fogg PC et al (2012) Comparative genomics of Shiga toxin encoding bacteriophages. BMC Genomics 13:311
39. Stothard P, Wishart DS (2005) Circular genome visualization and exploration using CGView. Bioinformatics 21(4):537–539
40. Grant JR, Arantes AS, Stothard P (2012) Comparing thousands of circular genomes using the CGView comparison tool. BMC Genomics 13: 202–213
41. Cresawn SG, Pope WH, Jacobs-Sera D et al (2015) Comparative genomics of cluster O mycobacteriophages. PLoS One 10(3): e0118725
42. Hatfull GF (2014) Molecular genetics of mycobacteriophages. Microbiol Spectr 2(2): 1–36
43. Pope WH, Bowman CA, Russell DA et al (2015) Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. elife 4: e06416
44. Bao Y, Chetvernin V, Tatusova T (2012) PAirwise sequence comparison (PASC) and its application in the classification of filoviruses. Viruses 4(8):1318–1327
45. Muhire BM, Varsani A, Martin DP (2014) SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One 9(9):e108277
46. Richter M, Rossello-Mora R, Oliver GF et al (2016) JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinformatics 32(6): 929–931
47. Goris J, Konstantinidis KT, Klappenbach JA et al (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57(Pt 1):81–91
48. Meier-Kolthoff JP, Klenk HP, Goker M (2014) Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. Int J Syst Evol Microbiol 64(Pt 2):352–356
49. Meier-Kolthoff JP, Goker M (2017) VICTOR: genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics 33(21): 3396–3404
50. Russell DA, Hatfull GF (2017) PhagesDB: the actinobacteriophage database. Bioinformatics 33(5):784–786
51. Hatfull GF, Jacobs-Sera D, Lawrence JG et al (2010) Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. J Mol Biol 397(1):119–143
52. Mavrich TN, Hatfull GF (2017) Bacteriophage evolution differs by host, lifestyle and genome. Nat Microbiol 2:17112
53. Grose JH, Casjens SR (2014) Understanding the enormous diversity of bacteriophages: the tailed phages that infect the bacterial family *Enterobacteriaceae*. Virology 468–470:421–443
54. Grose JH, Jensen GL, Burnett SH et al (2014) Genomic comparison of 93 *bacillus* phages reveals 12 clusters, 14 singletons and remarkable diversity. BMC Genomics 15:855
55. Zerbini FM, Siddell SG, Mushegian AR et al (2022) Differentiating between viruses and virus species by writing their names correctly. Arch Virol 167(4):1231–1234
56. Lefkowitz EJ, Dempsey DM, Hendrickson RC et al (2018) Virus taxonomy: the database of the international committee on taxonomy of viruses (ICTV). Nucleic Acids Res 46(D1): D708–Dd17
57. Fauquet CM, Mayo MA, Maniloff J et al (2005) VIIIth report of the international committee on taxonomy of viruses. Academic Press, London
58. Van Regenmortel MH (2003) Viruses are real, virus species are man-made, taxonomic constructions. Arch Virol 148(12):2481–2488
59. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410
60. Brister JR, Ako-Adjei D, Bao Y et al (2015) NCBI viral genomes resource. Nucl Acids Res 43(Database issue):D571–D5D7
61. O'Leary NA, Wright MW, Brister JR et al (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733–DD45
62. Sayers EW, Beck J, Brister JR et al (2020) Database resources of the National Center for biotechnology information. Nucleic Acids Res 48(D1):D9–d16
63. Turner D, Kropinski AM, Adriaenssens EM (2021) A roadmap for genome-based phage taxonomy. Viruses 13(3)

64. Gregory AC, Zayed AA, Conceição-Neto N et al (2019) Marine DNA viral macro- and microdiversity from pole to pole. Cell 177(5): 1109–23.e14
65. Bobay LM, Ochman H (2018) Biological species in the viral world. Proc Natl Acad Sci U S A 115(23):6040–6045
66. Moraru C, Varsani A, Kropinski AM (2020) VIRIDIC-A novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. Viruses 12(11)
67. Nishimura Y, Yoshida T, Kuronishi M et al (2017) ViPTree: the viral proteomic tree server. Bioinformatics 33(15):2379–2380
68. Simmonds P, Aiewsakun P (2018) Virus classification - where do you draw the line? Arch Virol 163(8):2037–2046
69. Aiewsakun P, Adriaenssens EM, Lavigne R et al (2018) Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. J General Virol 99(9):1331–1343
70. Bartlau N, Wichels A, Krohne G et al (2022) Highly diverse flavobacterial phages isolated from North Sea spring blooms. ISME J 16(2):555–568
71. Moraru C (2022) VirClust – a tool for hierarchical clustering, core gene detection and annotation of (prokaryotic) viruses. BioRxiv. https://doi.org/10.1101/2021.06.14.448304
72. Terzian P, Olo Ndela E, Galiez C et al (2021) PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform 3(3):lqab067
73. Cresawn SG, Bogel M, Day N, Jacobs-Sera D et al (2011) Phamerator: a bioinformatic tool for comparative bacteriophage genomics. BMC Bioinform 12:395
74. Lamine JG, DeJong RJ, Nelesen SM (2016) PhamDB: a web-based application for building Phamerator databases. Bioinformatics 32(13): 2026–2028
75. Turner D, Reynolds D, Seto D et al (2013) CoreGenes3.5: a webserver for the determination of core genes from sets of viral and small bacterial genomes. BMC Res Notes 6:140–146
76. Page AJ, Cummins CA, Hunt M et al (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31(22): 3691–3693
77. Bayliss SC, Thorpe HA, Coyle NM et al (2019) PIRATE: a fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. GigaScience 8(10)
78. Fouts DE, Brinkac L, Beck E et al (2012) PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res 40(22):e172
79. Anany H, Mahadevan P, Turner D et al (2022) ICTV report C. ICTV virus taxonomy profile: *Chaseviridae* 2022. J Gen Virol 103(4)
80. Liao YD, Tu J, Feng TY, Kuo TT (1986) Characterization of phage-Xp10-coded RNA polymerase. Eur J Biochem 157(3):571–577
81. Barylski J, Enault F, Dutilh BE et al (2020) Analysis of spounaviruses as a case study for the overdue reclassification of tailed phages. Syst Biol 69(1):110–123
82. Barylski J, Kropinski AM, Alikhan NF et al (2020) ICTV report C. ICTV virus taxonomy profile: *Herelleviridae*. J Gen Virol 101(4): 362–363
83. Committee ICoToVE (2020) The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. Nat Microbiol 5(5): 668–674
84. Krupovic M, Turner D, Morozova V et al (2021) Bacterial viruses subcommittee and archaeal viruses subcommittee of the ICTV: update of taxonomy changes in 2021. Arch Virol 166(11):3239–3244
85. Cook R, Brown N, Redgwell T et al (2021) INfrastructure for a PHAge REference database: identification of large-scale biases in the current collection of cultured phage genomes. PHAGE (New Rochelle) 2(4):214–223
86. Petrov VM, Ratnayaka S, Nolan JM, Miller ES, Karam JD (2010) Genomes of the T4-related bacteriophages as windows on microbial genome evolution. Virol J 7:292. https://doi.org/10.1186/1743-422X-7-292. PMID: 21029436; PMCID: PMC2993671
87. Kropinski AM (2003) Phage T1: a lambdoid phage with attitude? BEG News 18. http://www.biologyaspoetry.org/PDFs/bgnws018.pdf
88. Wolf YI, Silas S, Wang Y et al (2020) Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. Nat Microbiol 5(10):1262–1270
89. Staub NL, Poxleitner M, Braley A et al (2016) Scaling up: adapting a phage-hunting course to increase participation of first-year students in research. CBE Life Sci Educ 15(2)
90. Hatfull GF (2015) Innovations in undergraduate science education: going viral. J Virol 89(16):8111–8113

91. Jordan TC, Burnett SH, Carson S et al (2014) A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. MBio 5(1):e01051–e01013
92. Hanauer DI, Graham MJ, Betancur L et al (2017) An inclusive Research Education Community (iREC): impact of the SEA-PHAGES program on research outcomes and student learning. Proc Natl Acad Sci U S A 114(51): 13531–13536
93. Kleppen HP, Nes IF, Holo H (2012) Characterization of a *Leuconostoc* bacteriophage infecting flavor producers of cheese starter cultures. Appl Environ Microbiol 78(18): 6769–6772
94. Kot W, Hammer K, Neve H et al (2013) Identification of the receptor-binding protein in lytic *Leuconostoc pseudomesenteroides* bacteriophages. Appl Environ Microbiol 79(10): 3311–3314
95. Dereeper A, Guignon V, Blanc G et al (2008) Phylogeny.Fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res 36(Web Server issue):W465–W4W9
96. Letunic I, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49(W1):W293–W2w6
97. Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol 55(4): 539–552