IN HONOR OF MARC VAN REGENMORTEL

CrossMark

Virus classification – where do you draw the line?

Peter Simmonds[1] · Pakorn Aiewsakun[1,2]

Abstract

High-throughput sequencing (HTS) and its use in recovering and assembling novel virus sequences from environmental, human clinical, veterinary and plant samples has unearthed a vast new catalogue of viruses. Their classification, known by their sequences alone, sets a major challenge to traditional virus taxonomy, especially at the family and species levels, which have been historically based largely on descriptive taxon definitions. These typically entail some knowledge of their phenotypic properties, including replication strategies, virion structure and clinical and epidemiological features, such as host range, geographical distribution and disease outcomes. Little to no information on these attributes is available, however, for viruses identified in metagenomic datasets. If such viruses are to be included in virus taxonomy, their assignments will have to be guided largely or entirely by metrics of genetic relatedness. The immediate problem here is that the International Committee on Taxonomy of Viruses (ICTV), an organisation that authorises the taxonomic classification of viruses, provides little or no guidance on how similar or how divergent viruses must be in order to be considered members of new species or new families. We have recently developed a method for scoring genomic (dis)similarity between viruses (Genome Relationships Applied to Virus Taxonomy – GRAViTy) among the eukaryotic and prokaryotic viruses currently classified by the ICTV. At the family and genus levels, we found large-scale consistency between genetic relationships and their taxonomic assignments for eukaryotic viruses of all genome configurations and genome sizes. Family assignments of prokaryotic viruses have, however, been made at a quite different genetic level, and groupings currently classified as sub-families are a much better match to the eukaryotic virus family level. These findings support the ongoing reorganisation of bacteriophage taxonomy by the ICTV Phage Study Group. A rapid and objective means to explore metagenomic viral diversity and make evidence-based assignments for such viruses at each taxonomic layer is essential. Analysis of sequences by GRAViTy provides evidence that family (and genus) assignments of currently classified viruses are largely underpinned by genomic relatedness, and these features could serve as a guide towards an evidence-based classification of metagenomic viruses in the future.

The diversity and classification of viruses

Virus taxonomy is an essential element in the description of viruses and acts as a unified catalogue of their vast diversity and genetic interrelationships. Viruses are assigned in a hierarchy of taxonomic levels by the International Committee on Taxonomy of Viruses (ICTV; https://talk.ictvonline.org/). Viral diversity is, however, far greater than that of other organisms, with major differences in their genetic material (RNA or DNA) and configurations (double or single stranded), as well as the orientation of their encoded genes. Viral genomes may, furthermore, be distributed across several segments, sometimes packaged together in a virion, or often in separate virus particles, all of which are needed to infect a cell for replication to occur. Viral genomes come in various sizes, reflecting their diverse replication mechanisms and cellular interactions, as well as the varying structural complexity of their virions. The smallest virus genomes range from less than 2,000 bases, containing two genes, to 2.5 million base pairs, containing over 2,500 genes [1]. Similarly, there is extraordinary variability in virus particle morphology and size; some virus particles show icosahedral or more complex symmetry, while others may form filamentous, rectangular, bullet, or even

✉ Peter Simmonds
Peter.Simmonds@ndm.ox.ac.uk

Pakorn Aiewsakun
pakorn.aiewsakun@gmail.com

1 Nuffield Department of Medicine, University of Oxford, Peter Medawar Building, South Parks Road, Oxford OX1 3SY, UK

2 Department of Microbiology, Faculty of Science, Mahidol University, Bangkok 10400, Thailand

bottle-shaped nucleocapsids. Viruses infecting bacteria typically possess "tails" and "spikes", structural complexes that attach to and pierce the otherwise impermeable bacterial cell wall and injects viral DNA into the cytoplasm.

The taxonomy of cellular organisms, including microorganisms such as bacteria and unicellular fungi, is built on a common evolutionary framework – ultimately all eukaryotes share a last common ancestor, distinct from those of bacteria and archaea representing the other domains of life. These deeper relationships are largely recoverable from their genetic relationships; the phylogeny of core genes, such as those for ribosomal proteins, provides a reasonable representation of their evolutionary origin and divergence many billions of years ago. Unfortunately, the diversity of viruses prevents such a reconstruction of virus evolutionary histories – they lack any equivalent set of universally conserved genes on which to construct a phylogeny [9, 11, 19]. Viruses appear to have appeared on several occasions as parasitic companions of the various prokaryotic and eukaryotic life forms that they infect [13].

The current taxonomy of viruses has itself gradually evolved since the formation of what is now the ICTV in 1966. The vastly and rapidly expanding knowledge of virus diversity since that time and the advent of molecular methods for virus discovery and genetic characterisation have greatly increased the number of taxa assigned, with the current totals being 9 orders, 131 families, 46 subfamilies, 803 genera and 4,853 species, following the ratification vote after the 49th ICTV Executive Committee meeting [18]. Incorporating the hugely diverse collection of evolutionarily related (and perhaps unrelated) groups of viruses into a single, over-arching framework represents a considerable and ongoing organisational achievement, and the resources and descriptive catalogues of viruses (9th and 10th Report - https://talk.ictvonline.org/ictv-reports/) are vital resources underpinning the whole virus classification field.

Nevertheless, you might discover on closer inspection that the unified virus taxonomy is a rather ramshackle construction, with taxonomic assignment rules often being based on quite different and inconsistent criteria between virus groups. For example, the assignment of viruses to the order *Herpesvirales* is based on their morphology, and is independent of genomic relationships, which do not place their members into a coherent genetic group. Contrastingly, membership of the order *Tymovirales* is based upon possession of genetically related RNA polymerase genes, irrespective of the highly variable virion structures of their members.

Species-level classification

The species definition is particularly variable between virus groups. This taxonomic level was originally used as the primary division of viruses showing distinct disease patterns, epidemiology and host range. *Yellow fever virus*, *Carnation mottle virus* and *Salmon pancreas disease virus* are typical examples of literally hundreds of such disease-focussed designations. As nucleotide sequence information on classified species gradually accrued in the 1980s-1990s, it became clear that species were typically genetically distinct from each other, but there was no fixed sequence divergence threshold that defined members of the same and different species. As an example of what typifies large numbers of other species assignments throughout the ICTV taxonomy, members of different flavivirus species show pairwise distances in the RNA-dependent RNA polymerase (RdRp) gene of only 1.5% (between the species *Israel turkey meningoencephalomyelitis virus* and *Bagaza virus*), while members of *San Perlita virus* and *Ilheus virus* differ by 44%. As an example of even greater discrepancy, members of the species *Louping ill virus,* which infect grouse and sheep in upland Britain, lie entirely within the clade of the separate species *Tick-borne encephalitis virus*. The two were defined as separate species based on their marked differences in geographical range, host associations, and pathogenicity, and clearly not their genetic relationships in this case.

It has been entirely possible to maintain what is essentially a phenotypically based system of species assignments for many decades – these assignments do, after all, divide viruses into groups that differ in important medical, veterinary and agricultural properties that serve a major purpose for classification. Flexibility in what defines species is implicit in the polythetic species definition [27, 28] developed by Marc Van Regenmortel, in which constellations of properties, none of which would be individually essential for species inclusion or exclusion, are formulated to produce a highly intuitive and effective descriptive definition. Polythetic species definitions, indeed, have provided the basis for ICTV taxonomy assignments for nearly two decades. The difficulty that has arisen from this approach is that, increasingly, viruses are discovered by molecular methods and, most recently, by deep sequencing of environmental and other samples using high-throughput sequencing (HTS) methods. These techniques often simply provide a virus sequence but none of the typical sets of properties that might constitute a polythetic definition.

As HTS methods become increasingly used, and the catalogue of viruses known only by their sequence data continues to expand almost exponentially, genomics-based species assignments are clearly required to accommodate these into the ICTV taxonomy [10, 12, 21, 22, 31]. The difficulty is how to assign what often are purely sequence-based species thresholds in a manner that is not purely arbitrary and which is consistent with the taxonomy of other viruses whose phenotypic properties are known. Many opinions and proposals have been put forward on this crucial procedural issue over many years [7, 14, 29, 30], recently reviewed in reference
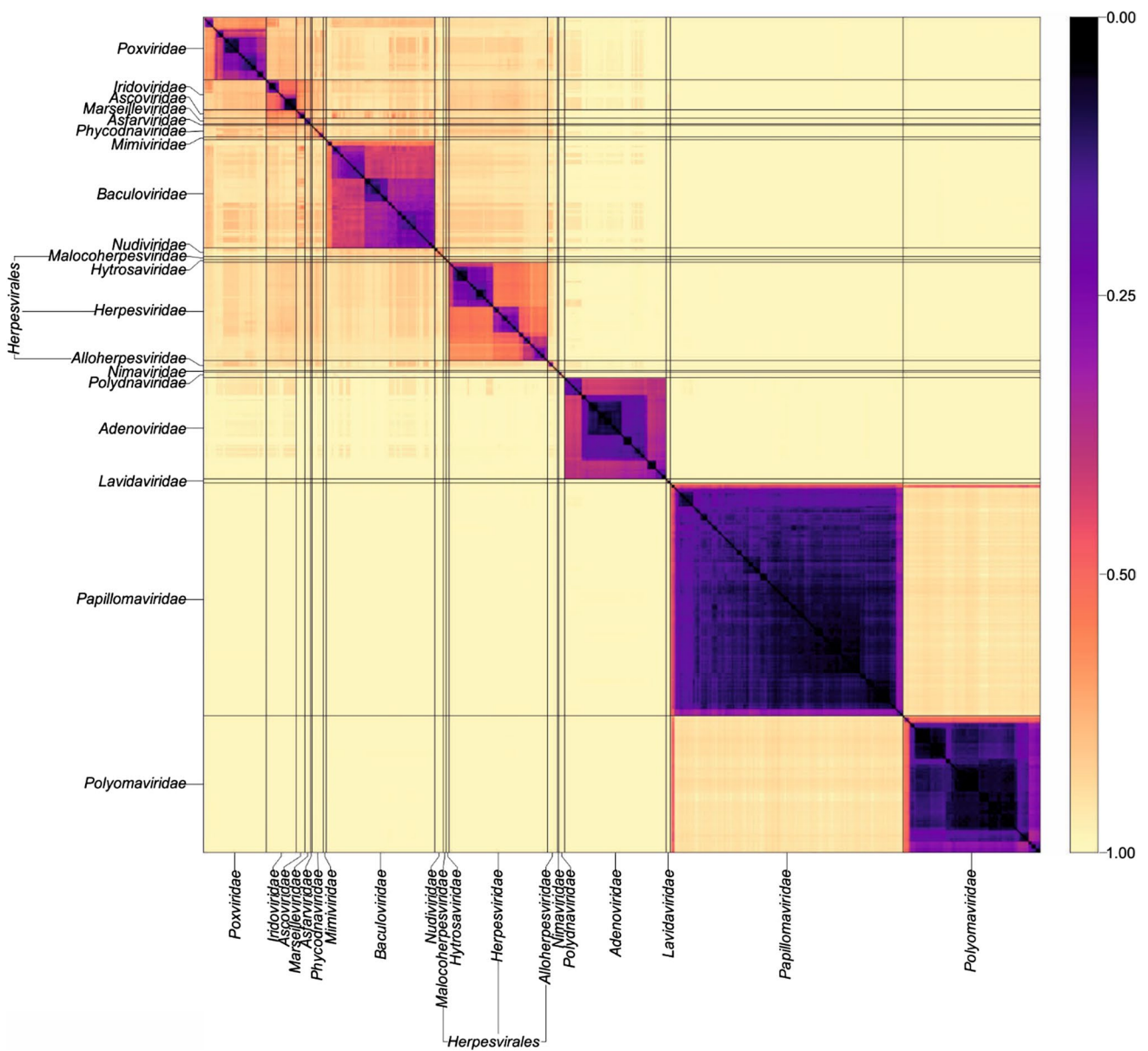
Fig. 1 Heat maps of CGJ distances between dsDNA. Pairwise CGJ distances were computed between each sequence plotted as a heat map using colour-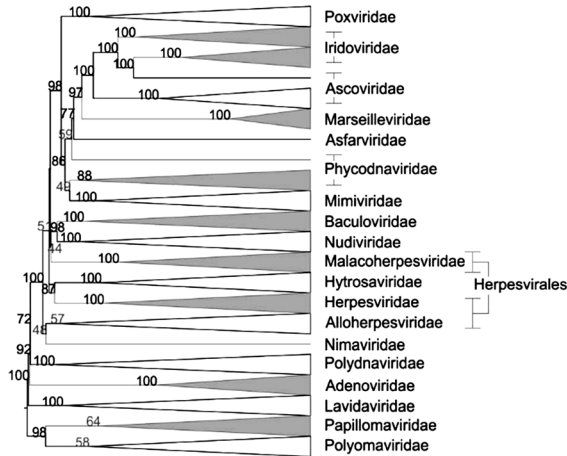coded points (see scale on the left of the figure). The light grey solid lines indicate boundaries between each virus taxonomic group. Annotations for each virus family and order are shown on the axes. This figure has been reproduced from reference [5]

[26]. However, in practical terms, all of these boil down to a central and seemingly irreconcilable dilemma – you can't use descriptive species definitions for viruses where there is nothing to describe except its nucleotide sequence. On the other hand, disallowing sequence-only assignments produces an ICTV taxonomy that fails in its function as a proper catalogue of viral diversity. Sequence-only viruses are still viruses. It is simply that information on their properties has not been collected (yet) – hardly a reason for their permanent exclusion. We suggest that those involved in previous discussions and those expressing new views on the subject
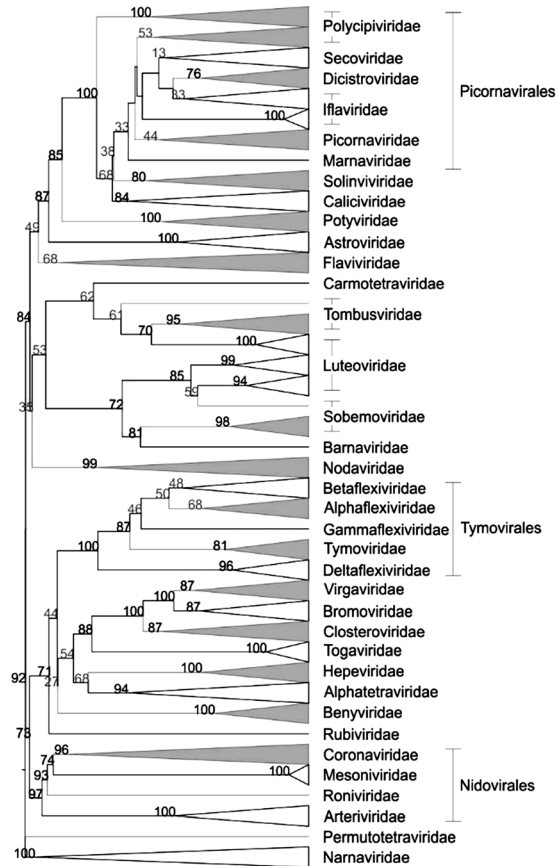
take the time to discuss future species definitions for viruses, to critically evaluate the various biological species definitions currently in use and to decide which concepts are most suitable for viruses in the future – the polythetic species is, after all, just one of the over 20-30 species definitions used or proposed in wider biology.

Virologists should not be daunted by the scale of the task ahead in making species assignment to the vast number of viruses identified by deep-sequencing methods. Recent descriptions of potentially tens or hundreds of new species, genus and family assignments for viruses isolated from
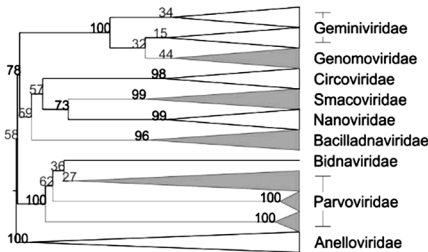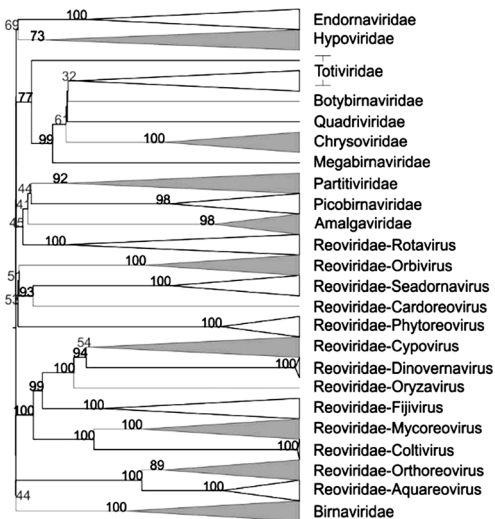
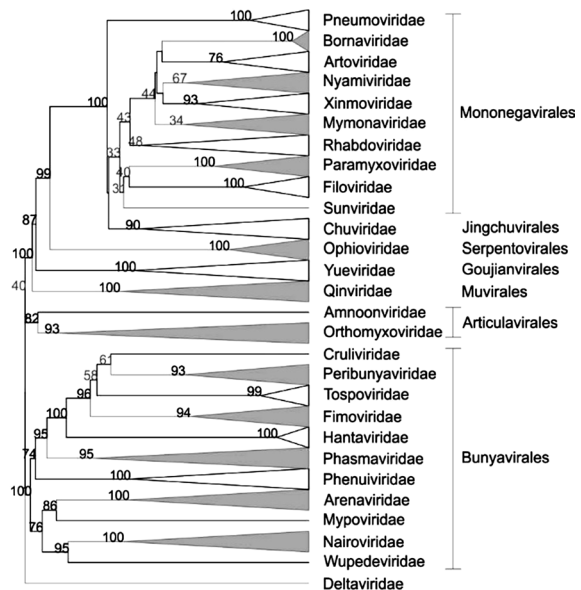## Group I: dsDNA viruses

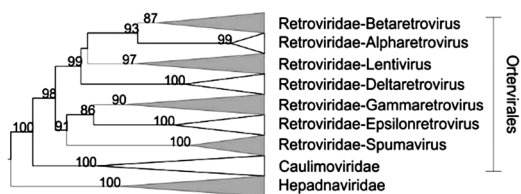## Group II: ssDNA viruses

Group III: dsRNA viruses

Group VI&VII: RT viruses

Group IV: (+)ssRNA viruses

Group V: (-)ssRNA viruses

arthropod and lower vertebrate host taxa [20, 23, 24] are really no different taxonomically from, and are indeed simpler in practical terms than classification tasks elsewhere in biology. A (possibly extreme) example is the still ongoing cataloguing of beetle species, with more than 400,000 species already assigned, each bearing a binomial, Latinised name and perhaps more than half a million to go.

Family-level classification

While the species level of taxonomy is one that best describes what we would consider individual viruses in a broad sense, the family level is the fundamental taxonomic level by which viruses of different kinds are organised. Eukaryotic viruses are currently assigned into a total of 111 families, with distinct genome configurations, virion morphologies, replication strategies and host interactions. Many of the groupings first described in early ICTV reports, generally based on electron microscopy appearance and genomic material, have stood the test of time – herpesviruses, poxviruses and rhabdoviruses are instantly and recognisably distinct morphologically, and their separateness has been verified by detailed analysis of their genome structures and replication mechanisms.

The ICTV taxonomy, however, provides little information that might guide decisions on what is needed to justify the creation of new virus families or orders. The ICTV Code provides this as a definition of a virus family:

> *A family is a group of genera (whether or not these are organized into subfamilies) sharing certain common characters*

building on:

> *A genus is a group of species sharing certain common characters*

which, frankly, is equally uninformative. Although new families have continued to be assigned in the last three decades, there is no real systematic information or guidance on what virological, structural or genomic attributes would support the family-level assignments that have been made, and what may be used in the future. For example, family members must presumably share homologous genes, but how many and how similar? Must they have related genome organisations? Should they look similar by electron microscopy? Most crucially in this "metagenomic" era, are there any family-defining attributes recoverable from a virus sequence alone – or is it essential to visualise particle morphology and to determine some of the physical attributes of the virus, such as host range, cell tropism and pathogenicity, as part of an extended family descriptive definition? The ICTV offers no guidance. Perhaps as a direct result of this uncertainty, picorna-like, flavi-like, circo-like and a vast range of other (family)–like virus descriptions are to be found in literally hundreds of papers indexed by PubMed. In many cases, it is as if their authors, too, lacked the information or confidence to assign new families for newly discovered viruses. Guidance on this issue and, ideally, a better-defined set of criteria for family assignments seems long overdue – this is indeed required if the ICTV is going to embrace the growing dataset of virus genome sequences from metagenomic datasets.

Development of GRAViTy

It was with this background that we set out to examine whether there were any consistent genomic attributes of eukaryotic virus families that correlated with their current family-level ICTV taxonomy relationships [5]. Investigating this involved extracting and analysing a wide range of genetic metrics from the set of currently classified viruses, including organisational features (gene complements and gene orders), possession of homologous genes and their amino acid sequence identity. These features were then systematically evaluated for their ability to recover the taxonomy of all currently classified eukaryotic viruses in the ICTV Master Species List [5]. The logic of this exercise was based on the premise that genome features that were informative and predictive of taxonomic relationships could then be extracted from currently unassigned viruses from their sequences alone (including the vast number of "-like" viruses), and informed decisions on their assignments could be made. While this represents an entirely bioinformatic approach, it ensures that any family or other-level assignments of sequence-only viruses are broadly consistent with existing taxonomic placements of viruses classified by other means.

This method, "Genome Relationships Applied to Virus Taxonomy" (GRAViTy), was applied to the complete list of 3,854 classified eukaryotic viruses with complete genome sequence data. Information such as gene contents, orientations and protein profiles were extracted from each genome sequence and systematically compared through the calculation of a composite generalised Jaccard (CGJ) distance, a
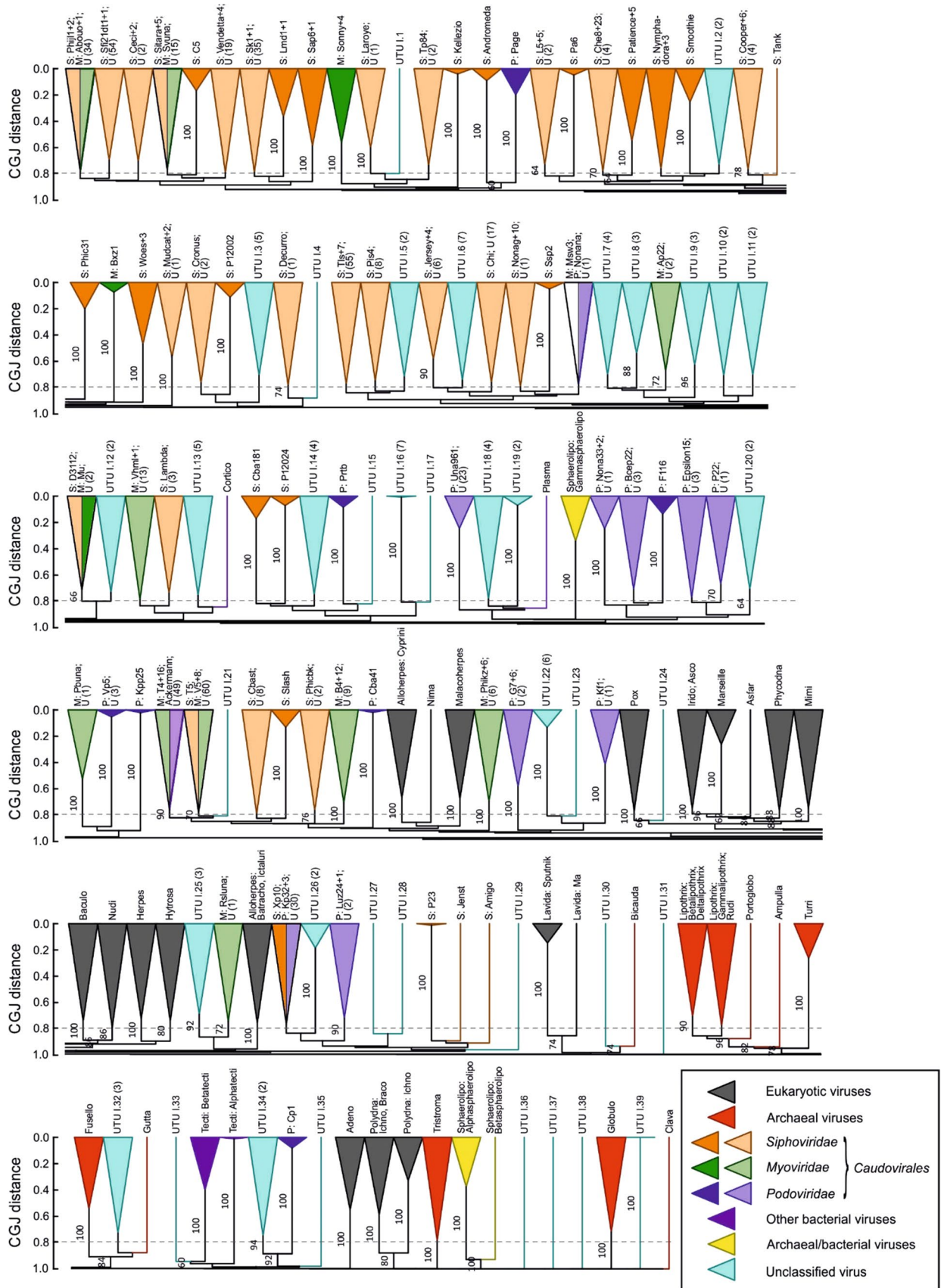
◄**Fig. 3** UPGMA dendrogram of classified and unclassified dsDNA viruses (Baltimore group I) based on CGJ distances. The dendrogram is divided into six separate lines to represent the 139 clades present in the dataset. Tips are labelled with genus for members of the *Caudovirales*; abbreviated as S: *Siphoviridae*, M: *Myoviridae*, and P: *Podoviridae*, and by family/genus for other bacterial, eukaryotic and archaeal viruses, or by accession number codes for unclassified viruses. The scale bar for CGJ distance is shown at the left of each line, and the 0.8 threshold that corresponds to eukaryotic family groupings is shown as a grey dotted line. Bootstrap re-sampling was performed with pruned signature tables as described previously [5]. Clades were coloured based on host origin according to the key; those containing both classified and unclassified (U) sequences are shown in a lighter shade. The 39 new candidate unassigned taxonomic units (UTUs) arising from the inclusion of current unclassified viruses are shaded in light blue. This figure has been reproduced from reference [4]

metric that captures and weights the contributions of different genome features between pairs of viruses. This use of multiple features extracted from viral sequences contrasts with traditional phylogenetic methods, in which virus relationships are often calculated from small, highly conserved portions of their genomes, such as the catalytic core of the RdRp gene.

Pairwise distances between members of particular virus groups, such as dsDNA viruses (Baltimore group I), for example, can be visualised as a heat map (Fig. 1) or as dendrograms (Fig. 2). The highly encouraging finding from this analysis was the close concordance between sequence groupings and their current ICTV assignments into families for viruses of all configurations (dsDNA, ssDNA, dsRNA, (+)ssRNA, (-)ssRNA and retro-transcribing viruses). GRAViTy showed 95.7%-100% (mean, 99.1%) sensitivity and 99.3%-100% (mean, 99.8%) specificity for the virus assignments with the current virus taxonomy. These associations are all the more striking for being derived from genome features without pre-selection for which ones might be considered to be informative, and furthermore being generated without having to construct multiple sequence alignments with its attendant but often hidden assumptions and problems. The analysis revealed a primary division of viruses at the family level that, with very few exceptions, was readily identifiable as tight clusters with ≥70% bootstrap support in the dendrogram (Fig. 2). The small number of exceptions in other Baltimore groups were in themselves often highly informative – the separation of rubella virus (genus *Rubivirus*) from the rest of the family *Togaviridae* concurs with a variety of other evidence that it should be re-classified into a separate family. Some virus families turned out to be far more diverse and often polyphyletic, including *Rhabdoviridae* and *Reoviridae,* and these represent future candidates for reorganisation of their taxonomy.

With this newly developed ability to predict family memberships from genomic sequences alone, we next analysed large datasets of virus sequences derived from metagenomic datasets to determine the extent to which they might be classified as new potential families, or as members of existing ones. Using the same metrics that differentiated virus families in each Baltimore group, we were able to provisionally assign viruses in this combined metagenomic dataset as four new families of ssDNA viruses, four dsRNA virus families, potentially as many as 101 new family-level groupings of (+)ssRNA viruses and 16 new (-)ssRNA families. Clearly, these results are preliminary, and conclusions about such a large number of new families require corroboration with other methods, including RdRp phylogenies in the case of future RNA viruses, as is the extension of classification to genus and species levels. What we can say, however, is that future assignments will certainly approximate to relationships that GRAViTy has shown. Furthermore, the new frameworks for each Baltimore group depict sets of virus relationships that are consistent with and extend virus taxonomy in a manner that effectively follows the rules used in the current classification.

Bacterial and archaeal virus (phage) taxonomy

Viruses infecting bacteria and archaea are extraordinarily abundant and diverse. Almost all of the currently classified bacterial viruses are, however, assigned to just three families of tailed phages, namely the *Myoviridae*, *Podoviridae* and *Siphoviridae* [2], a division based upon their distinct morphologies (myo-: long contractile, sipho-: long non-contractile and podo-: short tails). While it has been appreciated for many years that members of these families are highly diverse genetically, there has been little to no attempt to match these family assignment categories to those used for eukaryotic viruses. There is, consequently, no real idea of the extent to which bacteriophage assignments are taxonomically equivalent or not. This may not have mattered too much when different and largely independent scientific communities were engaged in their separate phage and eukaryotic virus research programmes, but the advent of metagenomic virus sequence data challenges this division. Bacterial, archaeal, and eukaryotic viruses abound in environmental samples that typically generate the largest metagenomic datasets – how can there be different rules for their classification when their hosts are not necessarily known?

We addressed this question directly through analysis of sequence datasets by GRAViTy that incorporated both eukaryotic and prokaryotic viruses in Baltimore groups I-IV (dsDNA, ssDNA, dsRNA and (+)ssRNA) [4]. This analysis revealed quite different sequence relationships, particularly between those of members of the tailed phage families *Siphoviridae, Myoviridae* and *Podoviridae* (assigned to the order *Caudovirales*) and those of eukaryotic viruses.
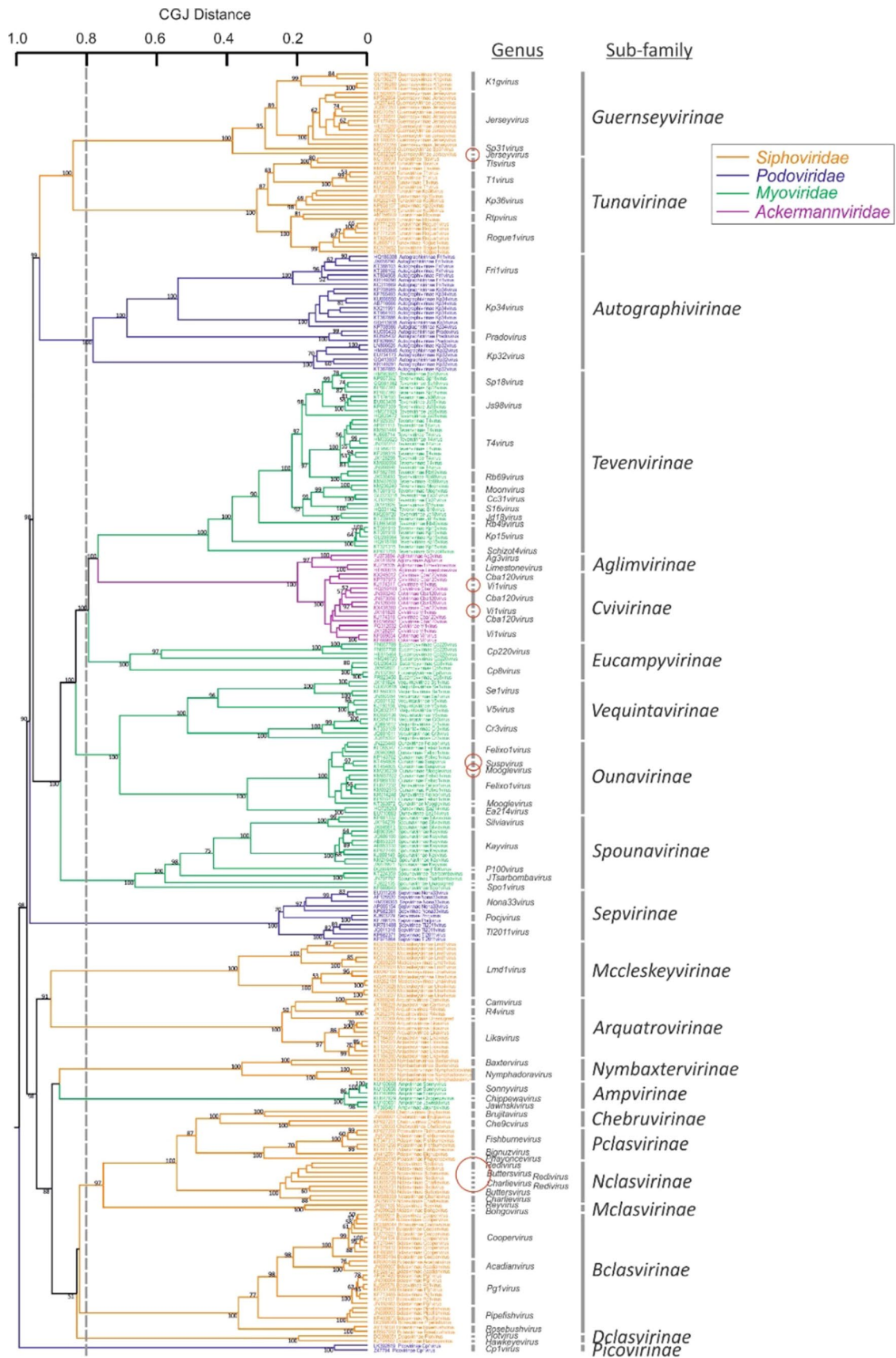
◄**Fig. 4** Dendrogram of members of the order *Caudovirales* that have subfamily assignments. Taxa are colour coded for family (see key). Minor discrepancies between genus assignments and phylogeny are shown in red circles. This figure has been reproduced from reference [4]

Members of the same phage family were generally far more divergent from each other and, despite their morphological similarities, frequently possessed genes that showed no detectable homology with each other (Fig. 3). If we were to apply the same metric of genetic relatedness observed between eukaryotic virus families, we might split the *Caudovirales* into as many as 70 family-level groupings and a further 37 groups comprising currently unclassified bacteriophage sequences available in public databases. These totals are, nevertheless, far more consistent with their own genetic diversity and that of the vast range of bacterial hosts they infect.

Examining sequence relationships a little further, it was apparent that the subfamily-level groupings among tailed phages were generally more consistent with family assignments of eukaryotic viruses where these have been made (Fig. 4). These findings fully support ongoing reclassifications by the ICTV of subfamilies, such as *Spounavirinae* and *Vi1virus* taxa, as new virus families based on genome relationships rather than phonotypic properties, such as virion appearance and presumed host [6, 15, 16]. The traditional morphology-based classifications of prokaryotic viruses are, indeed, increasingly untenable even in principle, since the bulk of new phage sequences to be classified derive from metagenomic datasets, where few or no phenotypic attributes are available.

Concluding remarks

The ICTV is committed to the incorporation of viruses known only by their genome sequence into the current taxonomy [3, 25]. This is a logical step if the ITCV taxonomy is to remain relevant as a guide to viral diversity; viral sequences reconstructed from HTS data represent viruses as much as those whose phenotypic properties are known (at least in part). However, as reviewed previously [26], classifications built on sequences alone directly challenge many of the concepts and assumptions of polythetic and other descriptive definitions of virus species.

They also reveal gaping holes in how the various taxonomic levels are defined – the ICTV code provides no indication of how different viruses have to be, either in terms of their phenotypic properties or in their genetic relatedness, to be considered members of the same or different species, genus or family. Given the importance of

the family level in virus taxonomy, it is remarkable that there is no information about what features might define a family and enable new families to be assigned in a consistent way. The application of bioinformatic programs like GRAViTy, vContact and others [4, 5, 8, 17] will be important in documenting those genomic features that underlie this and other taxonomic levels of currently classified eukaryotic and prokaryotic viruses. Such approaches are essential if future sequence-only assignments are to be made from a strong evidence base. While we can all foresee a future with a vastly expanded taxonomy of viruses, we think we should also preserve and respect as much of the current taxonomy as we can – the delineation of how current taxonomy works at a genomic level is an important initial step in this process.

Compliance with ethical standards

References

1. Abergel C, Legendre M, Claverie JM (2015) The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. FEMS Microbiol Rev 39:779–796
2. Ackermann HW, Prangishvili D (2012) Prokaryote viruses studied by electron microscopy. Arch Virol 157:1843–1849
3. Adams MJ, Lefkowitz EJ, King AM, Harrach B, Harrison RL, Knowles NJ, Kropinski AM, Krupovic M, Kuhn JH, Mushegian AR, Nibert ML, Sabanadzovic S, Sanfacon H, Siddell SG, Simmonds P, Varsani A, Zerbini FM, Orton RJ, Smith DB, Gorbalenya AE, Davison AJ (2017) 50 years of the International Committee on Taxonomy of Viruses: progress and prospects. Arch Virol 162:1441–1446
4. Aiewsakun P, Adriaenssens EM, Lavigne R, Kropinski AM, Simmonds P (2018) 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. https://doi.org/10.1099/jgv.0.001110
5. Aiewsakun P, Simmonds P (2018) The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. Microbiome 6:38

6. Barylski J, Enault F, Dutilh BE, Schuller MB, Edwards RA, Gillis A, Klumpp J, Knezevic P, Krupovic M, Kuhn JH, Lavigne R, Oksanen HM, Sullivan MB, Wittmann J, Tolstoy I, Brister JR, Kropinski AM, Adriaenssens EM (2018) Analysis of spounaviruses as a case study for the overdue reclassification of tailed bacteriophages. bioRxiv. https://doi.org/10.1101/220434

7. Blitvich BJ, Beaty BJ, Blair CD, Brault AC, Dobler G, Drebot MA, Haddow AD, Kramer LD, LaBeaud AD, Monath TP, Mossel EC, Plante K, Powers AM, Tesh RB, Turell MJ, Vasilakis N, Weaver SC (2018) Bunyavirus taxonomy: limitations and misconceptions associated with the current ICTV criteria used for species demarcation. Am J Trop Med Hyg 99(1):11–16

8. Bolduc B, Jang HB, Doulcier G, You ZQ, Roux S, Sullivan MB (2017) vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. PeerJ 5:e3243

9. Brussow H (2009) The not so universal tree of life or the place of viruses in the living world. Philos Trans R Soc Lond Ser B Biol Sci 364:2263–2274

10. Chow CE, Suttle CA (2015) Biogeography of viruses in the sea. Annu Rev Virol 2:41–66

11. Durzynska J, Gozdzicka-Jozefiak A (2015) Viruses and cells intertwined since the dawn of evolution. Virol J 12:169

12. Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3:504–510

13. Forterre P (2006) The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117:5–16

14. Gibbs AJ, Gibbs MJ (2006) A broader definition of 'the virus species'. Arch Virol 151:1419–1422

15. 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

16. Hulo C, Masson P, Le Mercier P, Toussaint A (2015) A structured annotation frame for the transposable phages: a new proposed family "Saltoviridae" within the Caudovirales. Virology 477:155–163

17. Iranzo J, Krupovic M, Koonin EV (2017) A network perspective on the virus world. Commun Integr Biol 10:e1296614

18. King AMQ, Lefkowitz EJ, Mushegian AR, Adams MJ, Dutilh BE, Gorbalenya AE, Harrach B, Harrison RL, Junglen S, Knowles NJ, Kropinski AM, Krupovic M, Kuhn JH, Nibert ML, Rubino L, Sabanadzovic S, Sanfacon H, Siddell SG, Simmonds P, Varsani A, Zerbini FM, Davison AJ (2018) Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018). Arch Virol. https://doi.org/10.1007/s00705-018-3847-1

19. Koonin EV, Senkevich TG, Dolja VV (2006) The ancient Virus World and evolution of cells. Biol Direct 1:29

20. Krupovic M, Ghabrial SA, Jiang D, Varsani A (2016) Genomoviridae: a new family of widespread single-stranded DNA viruses. Arch Virol 161(9):2633–2643

21. Mokili JL, Rohwer F, Dutilh BE (2012) Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2:63–77

22. Rosario K, Breitbart M (2011) Exploring the viral world through metagenomics. Curr Opin Virol 1:289–297

23. Shi M, Lin XD, Vasilakis N, Tian JH, Li CX, Chen LJ, Eastwood G, Diao XN, Chen MH, Chen X, Qin XC, Widen SG, Wood TG, Tesh RB, Xu J, Holmes EC, Zhang YZ (2015) Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the Flaviviridae and related viruses. J Virol 90:659–669

24. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS, Buchmann J, Wang W, Xu J, Holmes EC, Zhang YZ (2016) Redefining the invertebrate RNA virosphere. Nature 540:539–543

25. Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AM, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A, Zerbini FM (2017) Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15:161–168

26. Simmonds P (2018) A clash of ideas—the varying uses of the 'species' term in virology and their utility for classifying viruses in metagenomic datasets. J Gen Virol 99:277–287

27. Van Regenmortel MH, Maniloff J, Calisher C (1991) The concept of virus species. Arch Virol 120:313–314

28. Van Regenmortel MH (2003) Viruses are real, virus species are man-made, taxonomic constructions. Arch Virol 148:2481–2488

29. Van Regenmortel MHV (2016) Classes, taxa and categories in a heirarchical virus classification: a review of current debates of definitions and names of species. Bionomia 10(1):1–21

30. Van Regenmortel MHV (2018) The species problem in virology. Adv Virus Res 100:1–18

31. Wigington CH, Sonderegger D, Brussaard CPD, Buchan A, Finke JF, Fuhrman JA, Lennon JT, Middelboe M, Suttle CA, Stock C, Wilson WH, Wommack KE, Wilhelm SW, Weitz JS (2016) Re-examination of the relationship between marine virus and microbial cell abundances. Nat Microbiol 1:15024