frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 14 October 2022
DOI 10.3389/fmicb.2022.960465

OPEN ACCESS

EDITED BY
Robert Czajkowski,
University of Gdansk,
Poland

REVIEWED BY
Tao Jin,
Guangdong Magigene Biotechnology Co., Ltd, China
Jiatao Xie,
Huazhong Agricultural University, China

*CORRESPONDENCE
Li-Hong Yuan
ylh@gdpu.edu.cn
Jing-Zhe Jiang
jingzhejiang@gmail.com

SPECIALTY SECTION
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

RECEIVED 03 June 2022
ACCEPTED 30 September 2022
PUBLISHED 14 October 2022

CITATION
Yuan W-G, Liu G-F, Shi Y-H, Xie K-M, Jiang J-Z and Yuan L-H (2022) A discussion of RNA virus taxonomy based on the 2020 International Committee on Taxonomy of Viruses report.
Front. Microbiol. 13:960465.
doi: 10.3389/fmicb.2022.960465

A discussion of RNA virus taxonomy based on the 2020 International Committee on Taxonomy of Viruses report

Wen-Guang Yuan[1], Guang-Feng Liu[2], Ying-Hui Shi[1], Ke-Ming Xie[1], Jing-Zhe Jiang[1,2]* and Li-Hong Yuan[1]*

[1]Guangdong Province Key Laboratory for Biotechnology Drug Candidates, School of Biosciences and Biopharmaceutics, Guangdong Pharmaceutical University, Guangzhou, Guangdong, China, [2]Key Laboratory of South China Sea Fishery Resources Exploitation and Utilization, Ministry of Agriculture and Rural Affairs, South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, Guangdong, China

RNA viruses have a higher mutation rate than DNA viruses; however, RNA viruses are insufficiently studied outside disease settings. The International Committee on Taxonomy of Viruses (ICTV) is an organization set up by virologists to standardize virus classification. To better understand ICTV taxonomy and the characteristics and rules of different RNA virus families, we analyzed the 3,529 RNA viruses included in the 2020 ICTV report using five widely used metrics: length, host, GC content, number of predicted ORFs, and sequence similarity. The results show that host type has a significant influence on viral genome length and GC content. The genome lengths of virus members within the same genus are quite similar: 98.28% of the genome length differences within any particular genus are less than 20%. The species within those genera containing segmented viruses also have a similar length and number of segments. The number of predicted ORFs in the RNA viral genomes also shows a strong, statistically significant correlation with genome length. We suggest that due to the high mutation rate of RNA virus genomes, current RNA virus classification should mainly rely on protein similarities rather than nucleic acid similarities.

KEYWORDS

virus taxonomy, protein alignment, ICTV, RNA virus, segmented virus, Baltimore classification

Introduction

Virus classification is the process of naming viruses and placing them into a taxonomic hierarchy, as are the classification systems used for cellular organisms. On one hand, on the basis of virus host, viruses can be classified into four types, namely, animal viruses, fungi viruses, plant viruses, or bacteriophages (Bhat and Rao, 2020). On the other hand, to describe viruses more accurately, David Baltimore established a virus classification system

based on the manner of messenger RNA (mRNA) synthesis—the Baltimore classification system. This system classifies viruses into seven types: double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), double-stranded RNA (dsRNA), +strand single-stranded RNA (+ssRNA), −strand single-stranded RNA (−ssRNA), single-stranded RNA viruses with reverse transcriptase (ssRNA-RT), and double-stranded DNA viruses with reverse transcriptase (dsDNA-RT). The Baltimore classes remain an integral part of the conceptual foundation of biology (Koonin et al., 2021).

The International Committee on Taxonomy of Viruses (ICTV) was established in 1966 by virologists to standardize virus classification and naming. This established the first complex and complete virus classification system. In the newest ICTV taxonomy, RNA viruses (except ssRNA-RT) are classified into five major groups based on the phylogenetic tree constructed by Koonin et al. (Wolf et al., 2018). Their results show that dsRNA viruses evolved from +ssRNA viruses on at least two independent occasions, whereas −ssRNA viruses evolved from dsRNA viruses. Furthermore, the last common ancestors of the major branches of +ssRNA viruses only encode the RdRp (RNA-dependent RNA polymerase) and a single jelly-roll capsid protein in common with each other (Wolf et al., 2018).

RNA viruses (except ssRNA-RT) mutate rapidly with a mutation rate that is on average ≅2–3 orders of magnitude higher than DNA viruses. Even ssRNA-RT viruses have a mutation rate that is an order of magnitude higher than DNA viruses (Duffy et al., 2008). RNA virus nucleotide substitution rates are estimated to be roughly six orders of magnitude greater than those of corresponding cellular hosts (Holmes, 2009). These RNA virus characteristics inevitably increase the difficulty of classification. To better understand the most recent ICTV taxonomy and the characteristics and rules of different RNA virus families, we analyzed the 3,529 RNA viruses included in the 2020 ICTV report using five widely used metrics: length, host, GC content, number of predicted ORFs, and sequence similarity. Our review will provide support for analyzing the ICTV taxonomy and understanding the similarities and differences of different virus family members at the genome level.

Materials and methods

Downloading and construction of RNA virus database

We set up a localized RNA virus database for the 3,529 RNA viruses included in the 2020 ICTV report. Because of a lack genome data, we eventually downloaded only 2,249 nucleic acid sequences (Supplementary Table S1) from the National Center for Biotechnology Information (NCBI) on December 21, 2020.

Relationships between viral genome length, GC content, number of ORFs, and host

GC content and genome length were measured with the "seqkit" package (Wei et al., 2016) available in Linux. The Boxplot was drawn by R. Because the number of viruses infecting some hosts is insignificantly small and/or the complete genome is not available. We only counted those groups with complete genomes and in which the number of viruses infecting a type of host was >40, as illustrated in Figure 1. We annotated the viral genomes with Prodigal (Hyatt et al., 2010) and counted the resulting ORFs.

Analysis of differences among virus genus levels

To better analyze any possible correlations between the length of a virus genome and the genus level of the virus, we quantified the genus genome differences (equation 1).

$$\text{Genus genome differences} = \frac{(\text{Viral genome length}) - (\text{Average length of inter generic virus genomes})}{(\text{Average length of inter generic virus genomes})} \quad (1)$$

Sequence similarity analyses

The 2,249 RNA virus genome sequences were compared at the protein level using the tblastx function in BLAST (Altschul et al., 1990), and the maximum identity between any two sequences was taken as the protein similarity. Using a custom k-mer algorithm (Kirk et al., 2018) in this study (k-mer set to 10), the 2,249 RNA virus genome sequences were also compared at the nucleic acid level. The nucleic acid similarity between the sequences was calculated by equation (2).

$$\text{Similarity}_{\text{K-mer}} = 2^* \frac{(\text{Number of the same segments})}{(\text{Summer of the segments})} \quad (2)$$

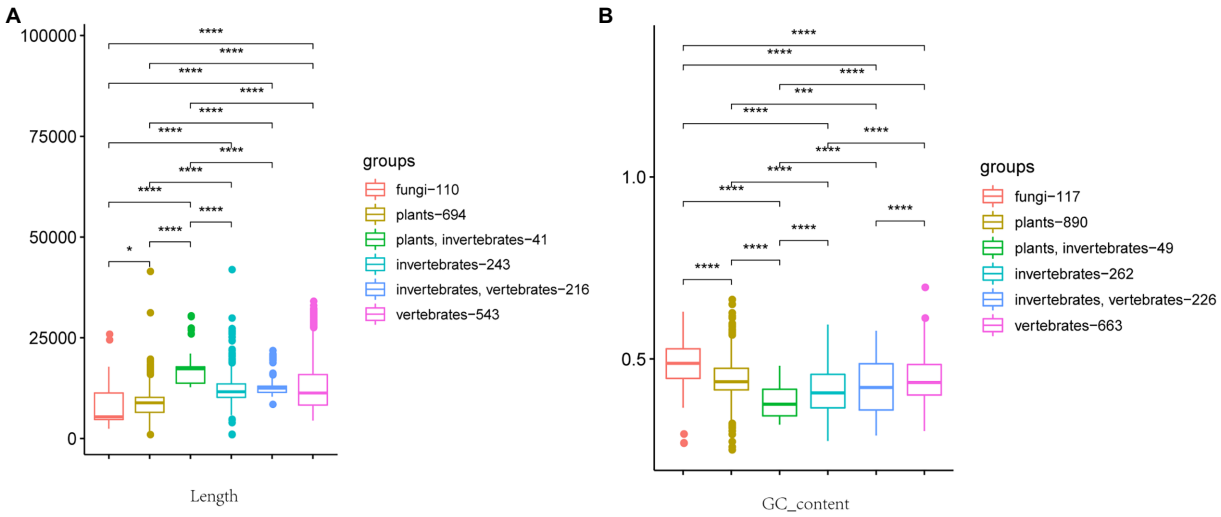

FIGURE 1
Box plot of relationship between viral genome length, GC content, and host. **(A)** Relationship between viral genome length and host. **(B)** Relationship between GC content and host. The host from left to right is fungi, plants, plants+invertebrates, invertebrates, invertebrates+vertebrates, and vertebrates. The number after the host represents the number of viruses included in the statistics. (Wilcoxon rank-sum test, *$p \leq 0.1$, ****$p \leq 0.0001$.)

TABLE 1 the 2020 RNA virus ICTV taxonomy.

Realm	Kingdom	Phylum
Riboviria (RNA + dsDNA-RT)	*Orthornavirae* (RDRP)	*Duplornaviricota* (dsRNA)
		Kitrinoviricota (+ssRNA)
		Lenarviricota (+ssRNA)
		Negarnaviricota (-ssRNA)
		Pisuviricota
	Pararnavirae (RT)	*Artverviricota*

Results

Adjustment of ICTV To RNA virus taxonomy

The ICTV Executive Committee approved the new taxonomic changes in August 2020 (Gorbalenya et al., 2020). In the new version of the taxonomy, they extended the previously established realm *Riboviria* to almost all RNA viruses and retroviruses (Walker et al., 2020). *Riboviria* is now divided into two kingdoms according to viral replication mode. One, *Orthornavirae*, uses RdRp to replicate. Following Wolf et al.'s phylogenetic tree, this kingdom is divided into five phyla (Wolf et al., 2018; Table 1): *Duplornaviricota*, *Kitrinoviricota*, *Lenarviricota*, *Negarnaviricota*, and *Pisuviricota*. The other *Riboviria* kingdom, *Pararnavirae*, uses RT for reverse transcriptional replication. In this kingdom, because of the low number of viruses found so far, only one phylum, *Artverviricota*, exists and is currently divided into six families (Krupovic et al., 2018): *Retroviridae*, *Metaviridae*, *Caulimoviridae*, *Belpaoviridae*, *Pseudoviridae*, and *Hepadnaviridae*. In summary, ICTV taxonomy is still mainly based on the Baltimore classification system. However, the classification has been adjusted and subdivided according to the phylogenetic tree proposed by Koonin et al. (Wolf et al., 2018).

Variation in genome length between hosts

We tabulated the length of 2,249 RNA virus genomes and associated those lengths with hosts. After preprocessing, six major categories were ultimately differentiated according to host type: fungi, plants, vertebrates, invertebrates, vertebrates + invertebrates, and plants + invertebrates. As shown in Figure 1A, RNA virus genome length generally shows significant differences between hosts. Vertebrate + invertebrate is more concentrated in the 12,000 bp category, indicating that the genome size of such cross-host transmissible viruses is closely related to the host. In addition, the average genome size of viruses associated with animal-associated virus groups (vertebrates, invertebrates, vertebrates + invertebrates, and plants + invertebrates) is significantly larger than that of the other two host types (plants and fungi).

Variation in GC content between hosts

We also tabulated the GC content of each virus genome sequence and associated that value with hosts. As with genome lengths, GC content generally shows significant differences between hosts (Figure 1B). Correspondingly the average GC

content of fungi viruses is significantly higher than that of other groups (Figure 1B; Hettiarachchi and Saitou, 2016). Because viruses need to mobilize the function of host cells to replicate and multiply when a virus infects its host, the GC content of a virus genome reflects the GC content of its host genome.

Differences in genome length and number of ORFs among different taxa

To visualize relationships between virus genome lengths and ICTV virus taxa, we tabulated the 2,035 viral genome lengths of those viruses with complete genomes and aligned those genomes at the order, family, and genus levels. As shown in Figure 2 (blue), viral genome lengths within the same family show good consistency. In particular, the lengths of virus genomes within genera are quite consistent. Except for 35 viral genomes with genus genome differences (equation (1)) are more than 20%, most of the rest (98.28%) of the genus genome differences are less than 20% (Supplementary Table S2)). This means that the length of a viral genome may be used as an important basis for the classification of RNA viruses at the genus level.

In a similar process to our analysis of genome lengths, we visualized relationships between the number of ORFs within the ICTV virus taxa. As shown in Figure 2 (green), the number of ORFs within genera is quite consistent. The number of predicted ORFs in DNA phage genomes shows a strong, statistically significant correlation with genome length (Ha and Denver, 2018). We also analyzed the relationship between the number of predicted ORFs and the lengths of RNA viruses. The number of predicted ORFs in RNA viral genomes again shows a strong, statistically significant correlation with genome length (Supplementary Figure S1).

Comparison of segmented and non-segmented viruses

We statistic the genome length of segmented RNA viruses (585 total) in two ways. One is the length of single segment (Single-segment), and the other is the length of total length of all segments of the virus (Multiple-segments). As shown in Figure 3, single-segment (red) viruses have two length peaks, while multiple-segments (blue) and non-segmented (green) viruses have relatively single length peaks. The average length of multiple-segment viral genomes is significantly larger than that of non-segmented viruses (Supplementary Figure S2). Given that long RNA virus genomes are unstable, this suggests that the multiple-segments approach of RNA viruses can better accommodate the instability of RNA genomes. Furthermore, multiple-segments viruses within a genus have similar lengths and numbers of segments (Supplementary Table S3). For

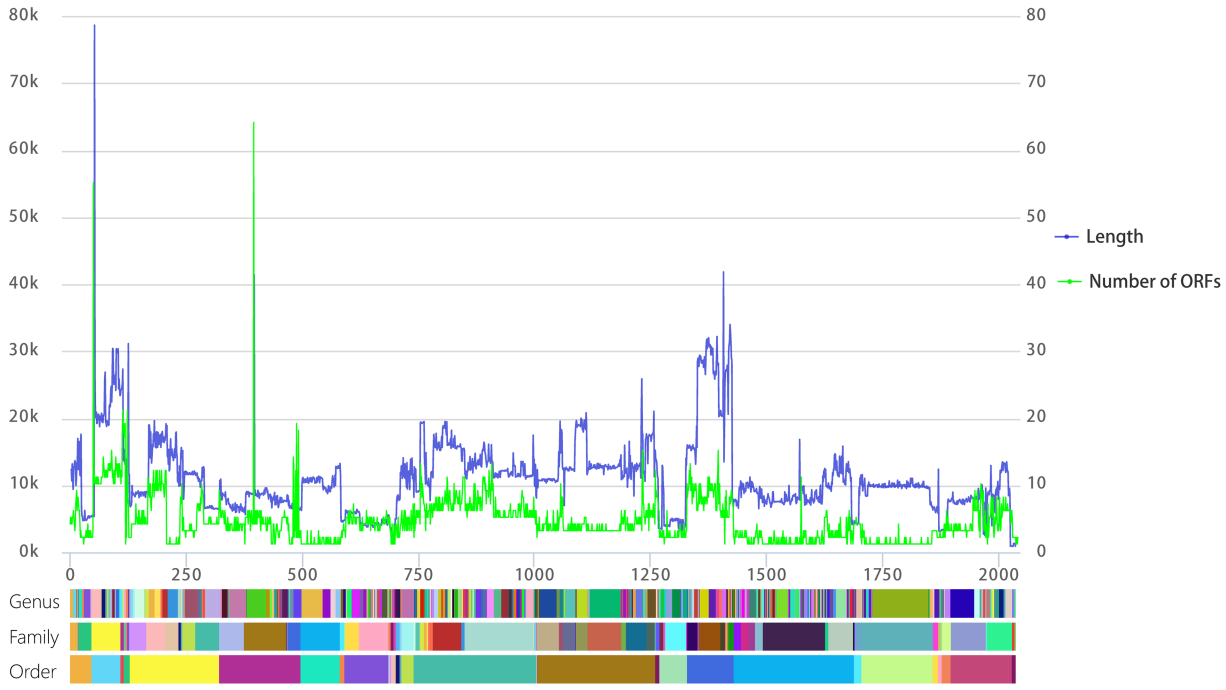

FIGURE. 2
Line chart of viral genome length and the number of ORFs aligned at the level of genus, family, and order. The left vertical axis is the length of the viral genome, the right vertical axis is the number of predicted ORFs, and the horizontal axis represents the cumulative number of viral genomes. The color blocks below correspond to the genus, family, and order from top to bottom, and each different color block represents a different genus, family, or order.

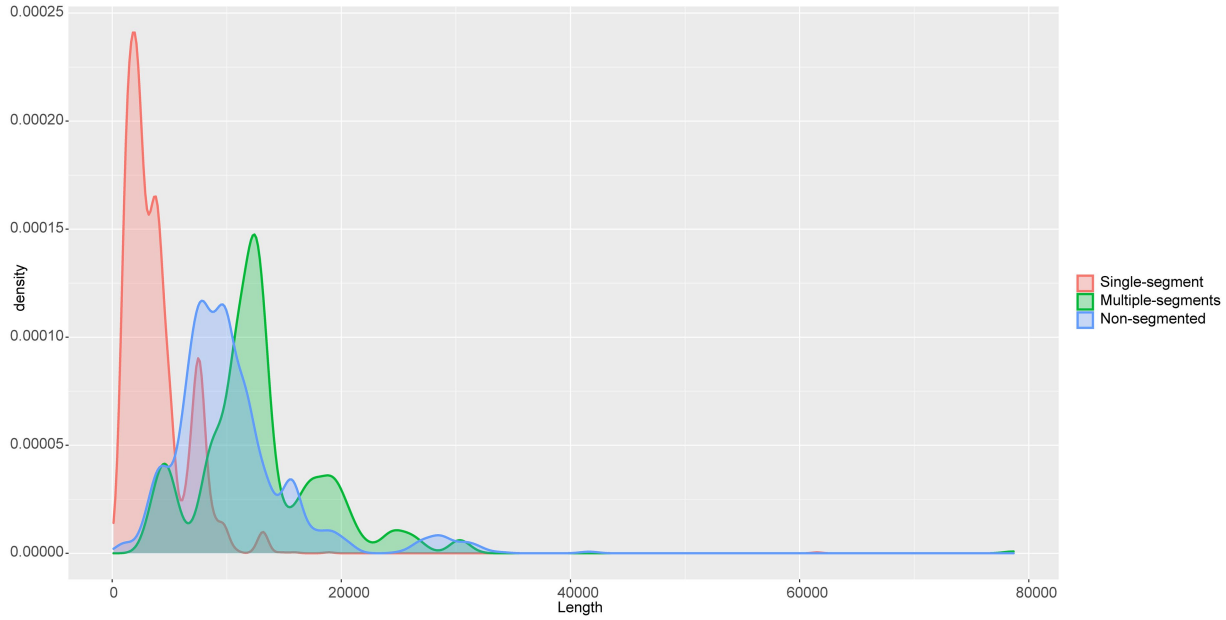

FIGURE. 3
Density plot of genome lengths between single-segment, multiple-segments, and non-segmented viruses. The horizontal axis is the length of the viral genome.

example, all viruses in genus *Furovirus* and *Mammarenavirus* are composed of one segment with a length of about 7,500 bp and another around 3,700 bp in length. All the viruses in the family *Bromoviridae* consist of one segment of about 3,500 bp in length and two other segments, each around 2,800 bp in length. It is precisely because of this consistency that the single-segment grouping has two peaks at 3,000 and 8,000 bp (Figure 3).

Nucleic acid and protein sequence similarity analyses

We performed pairwise similarity comparisons at the nucleic acid level (k-mer = 10) and protein level (tblastx) for all *Orthornavirae* and *Pararnavirae* genome sequences. As shown in Figure 4, similarity at the protein level in both virus kingdoms shows an obvious clustering effect. This indicates that viruses in the same family or genus have more similar protein sequences, as to be expected. However, similarity at the nucleic acid level does not show an obvious clustering effect. Therefore, due to the high mutation rate of RNA viruses, the classification of RNA viruses should be based on similarity at the protein level.

Discussion

Holmes described RNA virus evolution as being dominated by mutational processes. Because high error rates place an upper limit on genome size, it is extremely difficult to acquire the additional genetic material needed for a greatly improved polymerase (that is, one possessing some repair function) without suffering a mutational meltdown, as longer genomes result in a greater mutational burden (except in coronaviruses; Holmes, 2009). The results of our study show that the length of an RNA virus genome, whether it is segmented or non-segmented, shows strong regularity. This means that when an unknown RNA virus is taxonomically classified to a known viral family, its genome length should be an important reference factor in its classification.

Furthermore, based on our study, ICTV RNA virus taxonomy should be based on protein similarity rather than nucleic acid similarity. Given the high mutation rate of RNA virus genomes, it is more reasonable to use more conservative protein sequences to classify RNA viruses. This is the reason that most virus classification software, such as vConTACT2 (Bin Jang et al., 2019), CAT (von Meijenfeldt et al., 2019), and PhaGCN2,[1] are based on protein sequences. Furthermore, even the construction of RNA virus phylogenies should be based on protein sequence alignments. Tools based on nucleic acid sequences, such as Kraken2 (Wood et al., 2019), are limited to the identification of known virus sequences.

Consequently, the ICTV has changed its code to allow a 15-rank classification hierarchy (Gorbalenya et al., 2020). However, the ICTV classification method is still disputed.

1 https://github.com/KennthShang/PhaGCN2.0

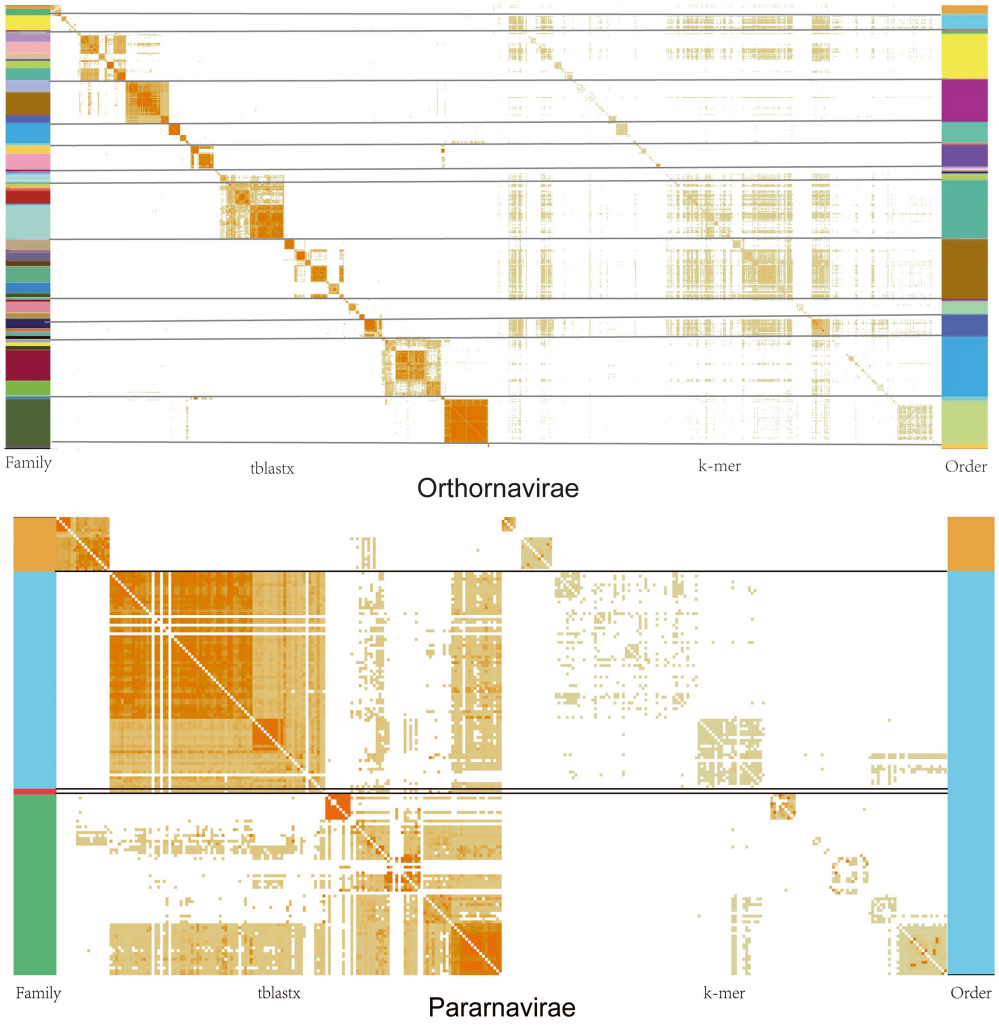

FIGURE. 4
Similarity matrices of viral genomes at the nucleic acid and protein levels in the *Orthornavirae* and *Pararnavirae*. The darker the orange in the figure, the greater the similarity. The different colors on the left represent the classification of different families, and the different colors on the right represent different orders.

Although the classification of RNA viruses at the level of family and order is considered valid (RESULT 3.3), it is not enough to classify an RNA virus at the phylum level according to Holmes and Duchene (2019). Presently, it is insufficient to rely solely on phylogeny to reconstruct the evolution of the global virome, but this is no reason to give up on global analyses of virus evolution (Wolf et al., 2019). At present, ICTV relies more on manual comparisons and phylogenetic analyses. However, with the discovery of more and more virus sequences (Wolf et al., 2020; He et al., 2022), current methods are not suitable for a large number of unknown virus classifications (Dutilh et al., 2021). Continuous development and testing of computational tools will be required to maintain a dynamic virus taxonomy that can accommodate new discoveries (Dutilh et al., 2021; Zayed Ahmed et al., 2022). However, the development of new classification tools, such as PhaGCN2 (footnote 1)—a semi-supervised machine learning model to classify viruses based on a graph convolutional network—may be a viable development direction.

Conclusion

We conducted a statistical analysis of the RNA virus genomes included in the 2020 ICTV report and found that host type has a significant impact on virus genome length, GC content, and segmentation. In particular, virus members within the same genus are more consistent in terms of genome length. Genomes length can be used as an important basis for RNA virus classification. This study also proposes that due to the high mutation rate of RNA virus genomes, the classification of RNA viruses should mainly rely on protein similarity rather than nucleic acid similarity.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

W-GY: methodology, validation, formal analysis, data curation, writing–original draft, and visualization. G-FL, Y-HS, and K-MX: resources, data curation, and investigation. J-ZJ: conceptualization, methodology, writing–original draft, writing–review and editing, supervision, project administration, and funding acquisition. L-HY: conceptualization, supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This project was supported by the Natural Science Foundation of China (nos. 31872499 and 31972847) to L-HY and J-ZJ; the Key-Area Research and Development Program of Guangdong Province (no.2022B1111030001) to J-ZJ; the Central Public-Interest Scientific Institution Basal Research Fund, CAFS (nos. 2020TD42 and 2021SD05) to J-ZJ; and the Guangdong Provincial Special Fund for Modern Agriculture Industry Technology Innovation Teams (no. 2019KJ141) to J-ZJ.

Acknowledgments

We thank Steven M. Thompson from Liwen Bianji (Edanz) (www.liwenbianji.cn/), for editing the English text of two drafts of this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.960465/full#supplementary-material

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Bhat, A. I., and Rao, G. P. (2020). "Host range of viruses," in *Characterization Plant Viruses : Methods and Protocols*. (New York, NY: Springer US), 29–31.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* 37, 632–639. doi: 10.1038/s41587-019-0100-8

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9, 267–276. doi: 10.1038/nrg2323

Dutilh, B. E., Varsani, A., Tong, Y., Simmonds, P., Sabanadzovic, S., Rubino, L., et al. (2021). Perspective on taxonomic classification of uncultivated viruses. *Curr. Opin. Virol.* 51, 207–215. doi: 10.1016/j.coviro.2021.10.011

Gorbalenya, A. E., Krupovic, M., Mushegian, A., Kropinski, A. M., Siddell, S. G., Varsani, A., et al. (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.* 5, 668–674. doi: 10.1038/s41564-020-0709-x

Ha, A. D., and Denver, D. R. (2018). Comparative genomic analysis of 130 bacteriophages infecting bacteria in the genus pseudomonas. *Front. Microbiol.* 9:1456. doi: 10.3389/fmicb.2018.01456

He, W. T., Hou, X., Zhao, J., Sun, J., He, H., Si, W., et al. (2022). Virome characterization of game animals in China reveals a spectrum of emerging pathogens. *Cell* 185, 1117–1129.e1118. doi: 10.1016/j.cell.2022.02.014

Hettiarachchi, N., and Saitou, N. (2016). GC content heterogeneity transition of conserved noncoding sequences occurred at the emergence of vertebrates. *Genome Biol. Evol.* 8, 3377–3392. doi: 10.1093/gbe/evw231

Holmes, E. C. (2009). The evolutionary genetics of emerging viruses. *Annu. Rev. Ecol. Evol. Syst.* 40, 353–372. doi: 10.1146/annurev.ecolsys.110308.120248

Holmes, E. C., and Duchene, S. (2019). Can sequence phylogenies safely infer the origin of the global Virome? *mBio* 10:e00289-19. doi: 10.1128/mBio.00289-19

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119

Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., et al. (2018). Functional classification of long non-coding RNAs by k-mer content. *Nat. Genet.* 50, 1474–1482. doi: 10.1038/s41588-018-0207-8

Koonin, E. V., Krupovic, M., and Agol, V. I. (2021). The Baltimore classification of viruses 50 years later: how does it stand in the light of virus evolution? *Microbiol. Mol. Biol. Rev.* 85:e0005321. doi: 10.1128/mmbr.00053-21

Krupovic, M., Blomberg, J., Coffin, J. M., Dasgupta, I., Fan, H., Geering, A. D., et al. (2018). Ortervirales: new virus order unifying five families of reverse-transcribing viruses. *J. Virol.* 92:e00515-00518. doi: 10.1128/JVI.00515-18

von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H., and Dutilh, B. E. (2019). Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* 20:217. doi: 10.1186/s13059-019-1817-x

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Dempsey, D. M., et al. (2020). Changes to virus taxonomy and the statutes ratified by the international committee on taxonomy of viruses (2020). *Arch. Virol.* 165, 2737–2748. doi: 10.1007/s00705-020-04752-x

Wei, S., Shuai, L., Yan, L., Fuquan, H., and Quan, Z. J. P. O. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* 11:e0163962. doi: 10.1371/journal.pone.0163962

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucia-Sanz, A., Kuhn, J. H., Krupovic, M., et al. (2018). Origins and evolution of the global RNA Virome. *mBio* 9:e02329-18. doi: 10.1128/mBio.02329-18

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucia-Sanz, A., Kuhn, J. H., Krupovic, M., et al. (2019). Reply to Holmes and Duchene, "can sequence phylogenies safely infer the origin of the global Virome?": deep phylogenetic analysis of RNA viruses is highly challenging but not meaningless. *mBio* 10:e00542-19. doi: 10.1128/mBio.00542-19

Wolf, Y. I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D., et al. (2020). Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat. Microbiol.* 5, 1262–1270. doi: 10.1038/s41564-020-0755-4

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biol.* 20:257. doi: 10.1186/s13059-019-1891-0

Zayed Ahmed, A., Wainaina James, M., Dominguez-Huerta, G., Pelletier, E., Guo, J., Mohssen, M., et al. (2022). Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* 376, 156–162. doi: 10.1126/science.abm5847