




# Can Sequence Phylogenies Safely Infer the Origin of the Global Virome?

 Edward C. Holmes,<sup>a,b,c</sup> Sebastián Duchêne<sup>d</sup>

<sup>a</sup>Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Sydney, NSW, Australia

<sup>b</sup>Charles Perkins Centre, School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW, Australia

<sup>c</sup>Sydney Medical School, The University of Sydney, Sydney, NSW, Australia

<sup>d</sup>Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, VIC, Australia

**KEYWORDS** evolution, phylogeny, RNA polymerase, virus

Resolving the origins of RNA viruses is one of the most important questions in viral evolution, but research on this question has been hampered by a lack of primary sequence similarity among the most divergent groups of RNA viruses (1). In a recent paper, Wolf et al. (2) seemingly overcame these problems and presented a comprehensive picture of the origins and deep evolutionary relationships among divergent RNA viruses. While many of the ideas presented in this study (2) were illuminating, we contend that they cannot be supported by the phylogenetic analysis performed, which is still based on sequences that often share no recognizable similarity and cannot be safely aligned.

Central to the study of Wolf et al. (2) is the quality of their sequence alignment of the most conserved RNA-dependent RNA polymerase (RdRp) protein. To obtain this, Wolf et al. (2) followed an iterative procedure, resulting in a main alignment comprising 4,627 taxa with a length of 12,220 amino acids. However, even a cursory inspection of this alignment indicates that it is highly unlikely to be accurate among the most distantly related RNA viruses. For example, (i) every site in the alignment contains at least one gap, including the canonical GDD motif, and there are no contiguous stretches of clearly aligned sequence across all viruses. (ii) Only 3.6% (441 residues) of the alignment remains after sites that harbor a majority (>50%) of gaps are removed. (iii) The pairwise identity between the aligned sequences is often less than the 5% expected by chance alone (mean value of 7.7% across the alignment) and was as low as 1%, with a mean pairwise distance of 0.93 (from a maximum of 1) substitutions per site. (iv) A total of 812 sites contain all 20 amino acids, and (v) 95.9% of sequences failed a  $\chi^2$  test of compositional heterogeneity in IQ-TREE (3). (vi) Only six sites can be safely aligned according to Gblocks (4), whereas TrimAl (5) could not align any sites, with both programs employing their least stringent settings. Points iii to v imply that even sophisticated substitution models cannot reliably estimate evolutionary divergence in these data (6), and point vi is particularly important for phylogenetic inference because the inclusion of poorly aligned regions results in biased tree estimates, with high bootstrap support for the incorrect topology (4). While clusters of clearly related sequences are present in these data, the deepest parts of the phylogeny reflect sequences that are so divergent in sequence that any attempt to depict their relationship, including through bootstrap analysis, is meaningless.

Despite the presence of a number of putative sequence motifs that we agree are indicative of common ancestry, the sequence alignment presented by Wolf et al. (2) is not sufficiently robust for a comprehensive phylogenetic analysis or to draw conclusions about the earliest moments of RNA virus evolution. We urge that caution be exercised in all studies that utilize sequences as divergent as those analyzed here, as

**Citation** Holmes EC, Duchêne S. 2019. Can sequence phylogenies safely infer the origin of the global virome? *mBio* 10:e00289-19. <https://doi.org/10.1128/mBio.00289-19>.

**Editor** Vincent R. Racaniello, Columbia University College of Physicians & Surgeons

**Copyright** © 2019 Holmes and Duchêne. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Edward C. Holmes, [edward.holmes@sydney.edu.au](mailto:edward.holmes@sydney.edu.au).

For the author reply, see <https://doi.org/10.1128/mBio.00542-19>.

**Published** 16 April 2019

phylogenies are meaningful only when they are estimated in the case of clear primary sequence similarity. Unfortunately, this is unlikely ever to be realistic in the case of RNA viruses.

## REFERENCES

1. Zotto PMDA, Gibbs MJ, Gould EA, Holmes EC. 1996. A reevaluation of the higher taxonomy of viruses based on RNA polymerases. *J Virol* 70:6083–6096.
2. Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, Dolja VV, Koonin EV. 2018. Origins and evolution of the global RNA virome. *mBio* 9:e02329-18. <https://doi.org/10.1128/mBio.02329-18>.
3. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol Biol Evol* 32:268–274. <https://doi.org/10.1093/molbev/msu300>.
4. Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 56:564–577. <https://doi.org/10.1080/10635150701472164>.
5. Capella-Gutiérrez S, Ailla-Martínez TM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973. <https://doi.org/10.1093/bioinformatics/btp348>.
6. Duchêne S, Di Giallonardo F, Holmes EC. 2016. Substitution model adequacy and assessing the reliability of estimates of virus evolutionary rates and time scales. *Mol Biol Evol* 33:255–267. <https://doi.org/10.1093/molbev/msv207>.