LETTER

Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable

Carla Mavian[a,b,1], Sergei Kosakovsky Pond[c], Simone Marini[a,d], Brittany Rife Magalis[a,b], Anne-Mieke Vandamme[e,f], Simon Dellicour[e,g], Samuel V. Scarpino[h], Charlotte Houldcroft[i], Julian Villabona-Arenas[j,k], Taylor K. Paisie[a,b], Nídia S. Trovão[l], Christina Boucher[m], Yun Zhang[n], Richard H. Scheuermann[n,o,p], Olivier Gascuel[q], Tommy Tsan-Yuk Lam[r], Marc A. Suchard[s,t,u], Ana Abecasis[f], Eduan Wilkinson[v], Tulio de Oliveira[v], Ana I. Bento[w], Heiko A. Schmidt[x], Darren Martin[y], James Hadfield[z], Nuno Faria[aa], Nathan D. Grubaugh[bb], Richard A. Neher[cc], Guy Baele[e], Philippe Lemey[e], Tanja Stadler[dd], Jan Albert[ee], Keith A. Crandall[ff], Thomas Leitner[gg], Alexandros Stamatakis[hh,ii], Mattia Prosperi[a,d,1], and Marco Salemi[a,b,1]

There is obvious interest in gaining insights into the epidemiology and evolution of the virus that has recently emerged in humans as the cause of the coronavirus disease 2019 (COVID-19) pandemic. The recent paper by Forster et al. (1) analyzed 160 severe acute respiratory syndrome coronavirus (SARS-CoV-2) full genomes available (https://www.gisaid.org/) in early March 2020. The central claim is the identification of three main SARS-CoV-2 types, named A, B, and C, circulating in different proportions among Europeans and Americans (types A and C) and East Asians (type B). According to a median-joining network analysis, variant A is proposed to be the ancestral type because it links to the sequence of a coronavirus from bats, used as an outgroup to trace the ancestral origin of the human strains. The authors further suggest that the "ancestral Wuhan B-type virus is immunologically or environmentally adapted to a large section of the East Asian population, and may need to mutate to overcome resistance outside East Asia." There are several serious flaws with their findings and interpretation. First, and most obviously, the sequence identity between SARS-CoV-2 and the bat virus is only 96.2%, implying that these viral genomes (which are nearly 30,000 nucleotides long) differ by more than 1,000 mutations. Such a distant outgroup is unlikely to provide a reliable root for the network. Yet, strangely, the branch to the bat virus, in figure 1 of their paper, is only 16 or 17 mutations in length. Indeed, the network seems to be misrooted, because

[a]Emerging Pathogens Institute, University of Florida, Gainesville, FL 32609; [b]Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, FL 32610; [c]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122; [d]Department of Epidemiology, University of Florida, Gainesville, FL 32610; [e]Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Clinical and Epidemiological Virology, Katholieke Universiteit Leuven, 3000 Leuven, Belgium; [f]Center for Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, 1349-008 Lisboa, Portugal; [g]Spatial Epidemiology Laboratory, Université Libre de Bruxelles, 1050 Bruxelles, Belgium; [h]Network Science Institute, Northeastern University, Boston, MA 02115; [i]Department of Medicine, University of Cambridge, Cambridge CB2 3QG, United Kingdom; [j]Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London WC1E 7HT, United Kingdom; [k]Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London WC1E 7HT, United Kingdom; [l]Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, MD 20892; [m]Department of Computer & Information Science & Engineering, University of Florida, Gainesville, FL 32611; [n]Department of Informatics, Craig Venter Institute, La Jolla, CA 92037; [o]Division of Vaccine Discovery, La Jolla Institute for Immunology, La Jolla, CA 92037; [p]Department of Pathology, University of California San Diego, La Jolla, CA 92093; [q]Unité Bioinformatique Evolutive, Centre de Bioinformatique, Biostatistique et BiologieIntégrative (C3BI)–USR 3756 CNRS and Institut Pasteur, 75015 Paris, France; [r]State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Hong Kong SAR, China; [s]Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095; [t]Department of Biostatistics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095; [u]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095; [v]KwaZulu-Natal Research Innovation and Sequencing Platform, University of KwaZulu-Natal, Durban 4041, South Africa; [w]Department of Epidemiology and Biostatistics, School of Public Health, Indiana University, Bloomington, IN 47405; [x]Max F. Perutz Laboratories, Center for Integrative Bioinformatics Vienna (CIBIV), Medical University of Vienna, University of Vienna, 1030 Vienna, Austria; [y]Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town 7925, South Africa; [z]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109; [aa]Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom; [bb]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT 06510; [cc]Biozentrum, University of Basel, 4056 Basel, Switzerland; [dd]Department of Biosystems Science and Engineering, Swiss Federal Institute of Technology, 4058 Basel, Switzerland; [ee]Department of Microbiology Tumor and Cell Biology, Karolinska Institutet, 17177 Stockholm, Sweden; [ff]Department of Biostatistics & Bioinformatics, Computational Biology Institute, Milken Institute School of Public Health, George Washington University, Washington, DC 20052; [gg]Theoretical Biology & Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545; [hh]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany; and [ii]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

Author contributions: C.M. and M.S. designed research; C.M., S.K.P., S.M., B.R.M., A.-M.V., S.D., S.V.S., C.H., J.V.-A., T.K.P., N.S.T., C.B., Y.Z., R.H.S., O.G., T.T.-Y.L., M.A.S., A.A., E.W., T.d.O., A.I.B., H.A.S., D.M., J.H., N.F., N.D.G., R.A.N., G.B., P.L., T.S., J.A., K.A.C., T.L., A.S., M.P., and M.S. analyzed data; and C.M., M.P., and M.S. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: salemi@pathology.ufl.edu, cmavian@ufl.edu, or m.prosperi@ufl.edu.

First published May 7, 2020.

(see their SI Appendix, figure S4) a virus from Wuhan from week 0 (24 December 2019) is portrayed as a descendant of a clade of viruses collected in weeks 1 to 9 (presumably from many places outside China), which makes no evolutionary (2) or epidemiological sense (3).

As for the finding of three main SARS-CoV-2 types, we must underline that finding different lineages in different countries and regions is expected with any RNA virus experiencing founder effects (2). According to Forster et al.'s (1) own analysis, a single synonymous mutation (nucleotide change in a gene that does not result in a modified protein) distinguishes type A from type B, while one nonsynonymous mutation (resulting in a protein with a single amino acid change) separates types A and C, and another one separates types B and C. Given SARS-CoV-2's fast evolutionary rate, random emergence of new mutations is entirely expected, even in a relatively short timeframe (4). When a viral strain is introduced and spreads in a new population, such random mutations can be propagated without them being selected or advantageous, due to founder effects. The fact that SARS-CoV-2 sequences show some geographical clustering is not new and is nicely and interactively shown on Nextstrain (5), but this cannot be used as a proof of biological differences unless backed by solid experimental data (6). This is particularly true for the work of Forster et al., since their findings are based on a nonrepresentative dataset of 160 genomes, with no significant correlation between prevalence of confirmed cases and number of sequenced strains per country (7, 8). The essential role of representative sampling is well documented in the literature (9), but was not acknowledged by the authors, who, instead, claim that their "network faithfully traces routes of infections for documented [COVID-19] cases," without taking into consideration missing viral diversity, or evaluating multiple transmission hypotheses that would be consistent with sequence data, or even providing any support on the robustness of the branching pattern in their network. Ultimately, no firm conclusion should be drawn without evaluating the probability of alternative dissemination routes.

The inappropriate application and interpretation of phylogenetic methods to analyze limited and unevenly sampled datasets begs for restraint about origin, directionality, and early clade/lineage inference of SARS-CoV-2. We feel the urgency to reframe the current debate in more rigorous scientific terms, given the dangerous implications of misunderstanding the true dispersal dynamics of SARS-CoV-2 and the COVID-19 pandemic.

Acknowledgments

We are grateful to Paul Sharp, Andrew Rambaut, and Nicola De Maio for their insightful suggestions and critical reading of our manuscript. C.M., B.R.M., M.P., and M.S. were, in part, supported by NSF Division of Environmental Biology Award 2028221 and NIH National Institute of Allergy and Infectious Diseases Award 1R21AI138815-01. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

1 P. Forster, L. Forster, C. Renfrew, M. Forster, Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9241–9243 (2020).
2 M. Salemi, A.-M. Vandamme, P. Lemey, *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* (Cambridge University Press, Cambridge, United Kingdom, 2009).
3 D. S. Hui *et al.*, The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. *Int. J. Infect. Dis.* **91**, 264–266 (2020).
4 O. G. Pybus, A. Rambaut, Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550 (2009).
5 J. Hadfield *et al.*, Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
6 N. D. Grubaugh, M. E. Petrone, E. C. Holmes, We shouldn't worry when a virus mutates during disease outbreaks. *Nat. Microbiol.* **5**, 529–530 (2020).
7 C. Mavian, S. Marini, M. Prosperi, M. Salemi, A snapshot of SARS-CoV-2 genome availability up to 30th March, 2020 and its implications. *bioRxiv*:10.1101/2020.04.01.020594. (2020).
8 S. Weaver, State of GISAID COVID-19 sequence availability (2020) https://observablehq.com/@stevenweaver/case-vs-sequence-count. Accessed 12 April 2020.
9 S. D. Frost *et al.*, Eight challenges in phylodynamic inference. *Epidemics* **10**, 88–92 (2015).