HHS Public Access

Author manuscript

IEEE Int Conf Bioinform Biomed Workshops. Author manuscript; available in PMC 2023 February 10.

Published in final edited form as:

IEEE Int Conf Bioinform Biomed Workshops. 2022 December ; 2022: 2940–2944. doi:10.1109/bibm55620.2022.9995364.

Transmission cluster characteristics of global, regional, and lineage-specific SARS-CoV-2 phylogenies

Mattia Prosperi,
Department of Epidemiology, College of Public Health and Health Professions, University of Florida Gainesville, Fl, USA

Brittany Rife,
Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, University of Florida Gainesville, Fl, USA

Simone Marini,
Department of Epidemiology, College of Public Health and Health Professions, University of Florida Gainesville, Fl, USA

Marco Salemi
Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, University of Florida Gainesville, Fl, USA

Abstract

The SARS-CoV-2 pandemic has been presenting in periodic waves and multiple variants, of which some dominated over time with increased transmissibility. SARS-CoV-2 is still adapting in the human population, thus it is crucial to understand its evolutionary patterns and dynamics ahead of time. In this work, we analyzed transmission clusters and topology of SARS-CoV-2 phylogenies at the global, regional (North America) and clade-specific (Delta and Omicron) epidemic scales. We used the Nextstrain's nCov open global all-time phylogeny (September 2022, 2,698 strains, 2,243 for North America, 499 for Delta21A, and 543 for Omicron20M), with Nextstrain's clade annotation and Pango lineages. Transmission clusters were identified using Phylopart, DYNAMITE, and several tree imbalance measures were calculated, including staircase-ness, Sackin and Colless index. We found that the phylogenetic clustering profiles of the global epidemic have highest diversification at a distance threshold of 3% (divergence of 10, where the tree sampled median is 49). Phylopart and DYNAMITE clusters moderately-to-highly agree with the Pango nomenclature and the Nextstrain's clade. At the regional and clade-specific scale, transmission clustering profiles tend to flatten and similar clusters are found at distance thresholds between 0.05% and 25%. All the considered phylogenies exhibit high tree imbalance with respect to what expected in random phylogenies, suggesting short infection times and antigenic drift, perhaps due to progressive transition from innate to adaptive immunity in the population.

Keywords

SARS-CoV-2; phylogenetics; transmission cluster

m.prosperi@ufl.edu .

I. Introduction

The global epidemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), causative agent of the coronavirus disease 2019 (COVID-19), has been ongoing since the end of 2019 –declared as pandemic in March 2020– to present days (September 2022), with over 600 million cases and over 6 million deaths recorded in the world (https://covid19.who.int/). Vaccines have been made available for emergency use as early as June 2020 in China, and August 2020 in Russia. In Europe and North America, vaccines received emergency authorization at the end of 2020. As of September 2022, over 12 billion vaccine doses have been administered worldwide.

Across three years, SARS-CoV global and regional spread presented in periodic waves, the virus diversified into different variants, each characterized by a constellation of mutations, with convergent emergence in some cases due to selective pressure from human host immunity [1]. The SARS-CoV-2 evolution during the pandemic seems characterized primarily by purifying selection, but a small set of sites appear to evolve under positive selection [2]. Virus variants have been associated with increased transmissibility, virulence and changes to antigenicity [3]-[5]. Over time, certain variants became dominant by replacing others, e.g., Alpha, Delta, Gamma, and the most recent Omicron. There is also evidence that some circulating SARS-CoV-2 lineages are recombinant [6]. Lineage and variant nomenclature is based on phylogenetic divergence among isolates as well as epidemiological evidence. A systematic definition of lineages, called Pango, has been introduced [7]. The World Health Organization developed a nomenclature for variants of concerns (https://www.who.int/activities/tracking-SARS-CoV-2-variants) and appointed a technical advisory group on virus evolution to develop an early warning system of variant emergence [8].

The enormous collaborative efforts of the scientific community produced very large data SARS-CoV-2 repositories, e.g., GISAID (https://gisaid.org/) with over 13 million sequence submissions, and insightful software tools to view the epidemic in real time, e.g., Nextstrain [9]. Nextstrain utilizes its own clade nomenclature (https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming). Dellicour *et al.* introduced a phylodynamic tool to analyze how dispersal dynamics of lineages could be affected by interventions, e.g., lockdowns [10]. At present, there is a variety of software tools specific for SARS-CoV-2 from sequence analysis to protein structure and interactome prediction [11].

The number of molecular epidemiology studies that analyzed the SARS-CoV-2 pandemic is staggering, and many of them helped shedding light on evolutionary dynamics of the virus [12], [13]. Analyses at the regional level helped identifying introduction events, variant spread, and possible effects of public health measures [14]-[18].

SARS-CoV-2 is still adapting in the human population, thus it is crucial to understand its evolutionary patterns and dynamics ahead of time. In this work, we aimed at probing the following questions: (1) Are the current clade and lineage nomenclatures consistent with intra-inter lineage phylogenetic diversity? (2) Are there differences in transmission

cluster shapes among regional epidemics and do they differ between variants? (3) Are the phylogenies imbalanced and do they suggest short infection times with antigenic drift, partial cross-immunity, with differences among regional or clade-specific trees? Accordingly, we analyzed both global and regional phylogenies of SARS-CoV-2, comparing the statistical phylogenetic clustering method Phylopart with the Nextstrain's clade and Pango lineage nomenclature, assessing how the transmission clusters change from the global, to the regional, to the variant-specific scale, and quantifying several tree imbalance measures with respect to random phylogenies and other known viruses.

II. Methods

We used Nextstrain's dashboard (https://nextstrain.org/) to select the nCov (i.e., SARS-CoV-2) open global all-time phylogeny scaled in genetic divergence, and then to sub-select the North American, Delta21A, and Omicron20M trees. Nextstrain phylogenies use sequence data and metadata from NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/), under the Open Data principles (https://opendatahandbook.org/guide/en/what-is-open-data/). This work does not use original data, but only Nexstrain's trees, with annotated geography and lineage. For replication purposes, and to acknowledge the data sources (including the effort of researchers who contribute to open data), we uploaded the Genbank accession numbers as well as all metadata to credit all data generators and authors at: https://github.com/DataIntellSystLab/phylopart-sarscov2.

We made two important assumptions in this study: (1) the sampling was representative and uniform; and (2) the tree inferred topologies were correct. Transmission clusters were identified using Phylopart [19] on mid-point rooted trees at multiple distance thresholds over a grid (from $5{\cdot}10^{-5}$ to 0.25 with 25 steps), estimating the overall patristic distance distribution through sampling of 1,000,000 tip pairs. We also used, DYNAMITE, a refined version of Phylopart that also considers internal nodes into clustering, in a 'dynamic' perspective [20]. Agreement among cluster sets was assessed using the adjusted Rand index [21]. All other analyses were performed using R (https://www.r-project.org/), using libraries: ape, mcclust, phytools, and treebalance. The following tree imbalance measures were considered: area per pair index; average leaf depth index; cherry index; Colless index; maximal difference in widths; Rogers index; rooted quartet index; Sackin index; staircase-ness; total cophenetic index [22]. We focused in particular on the staircase-ness [23] that, for every bifurcating tree node, quantifies if one branching contains more tips than the other one. Of note, we used multiple measures since a single one is usually not enough to discriminate imbalance [24], [25]. All indices were compared to distributions obtained from 200 randomly generated trees with the same number of tips as the real phylogenies.

III. Results

Nextstrain's September 2022 nCov open global all-time phylogeny included 2,698 strains. The regional North American, Delta21A, Omicron20M trees included 2,243, 499, and 543 strains, respectively.

The distribution of nodes at each tree level (from the root) in the global phylogeny deviated from what expected in random trees with the same number of tips, exhibiting slightly more nodes than expected towards the root and the tips of the tree, and significantly less in the central part. This could indicate either a problem with sampling, or different rates, likely from lineages emerging not uniformly from the tree. The median root-to-tip distance increased moderately (0.44 x tree level + 37, ρ=0.73), whilst the median branch length remained constant (Fig. 1).

Phylopart identified a clear peak of maximal cluster diversification for both the global and regional North American trees at the 3rd percentile of the overall distance distribution, i.e., a divergence of 10 where the median tree divergence was 49. For the Delta21A and Omicron20M the highest number of clusters was attained at the 4th percentile, but there was no clear peak across thresholds (Fig. 2).

The maximal agreement between Phylopart clustering and Pango lineages was found at a diversity threshold of 8th percentile, i.e., a divergence of 17, which was higher than the maximal cluster diversification threshold. The adjusted Rand index ranged between 0.48 and 0.55 (considering all tree tips or removing unclustered/singleton isolates, respectively). The maximal agreement between Phylopart and Nextstrain's clade was found at thresholds between 8th to 12th percentile (i.e., diverge of 17 to 19) and the adjusted Rand index ranged between 0.50 and 0.73. Fig. 3 shows the global tree where tips have been colored according to Phylopart's clustering (3rd and 8th distance percentile), Nextstrain's clade, and Pango lineage. Major discrepancies between Phylopart and Nextstrain's clade were found in 20B, Delta21J, Epsilon20C, 20A, 20D, Omicron21L, 20G, and Delta21A (agreement between 71.2% and 97.1%). The highest concordance was with Omicron22A, Omicron22B, Omicron22C, Omicron21K, Delta21I, Lambda21G, Eta21D, Theta21E, Kappa21B, EU120E, 20F, 19A, and 19B (agreement >99.7%).

When performing clustering with DYNAMITE, results were similar to those obtained with Phylopart. The best agreement with both Pango lineages and Nextstrain's clades was found at the 25th percentile threshold (ARI=0.72 for both Pango and clades). At the 10th percentile it was similar to Phylopart (ARI=0.41 and 0.46, respectively). Of note, by calculating the DYNAMITE's threshold using only the portion of the tree from the root to the current nodes being considered for clustering, i.e., in a truly dynamic way that ignores 'future' nodes of the tree, the ARI with Pango and clade increased to 0.8. For instance, the SC2 lineage classification was improved by allowing cluster assignment to depend only on what was happening with the outbreak thus far and not on the outbreak/tree in its entirety (i.e., retrospectively).

All the considered phylogenies exhibited high tree imbalance with respect to what expected in random phylogenies. The global tree presented the highest deviation from the random tree distribution –in terms of t-value– for the area per pair index, cherry index, Rogers index, and staircase-ness. The North American tree exhibited the highest deviation for the maximal difference in widths. The Delta21A had the highest deviation for total cophenetic coefficient, Colless index, Sackin index, root quartet index, and average leaf depth index. Of note, every single t-value yielded a p-value below 0.0001, even after adjustment for multiple testing.

IV. Discussion

The transmission cluster characteristics of multi-lineage global and regional SARS-CoV-2 phylogenies are more similar than lineage-specific phylogenies. The clustering of the global tree as estimated by Phylopart is moderately-to-highly correlated with the Nexstrain's clade and Pango nomenclature, and similar, yet better in terms of ARI, results are obtained when using DYNAMITE. Although lineages and clades incorporate epidemiological evidence besides mere phylogenetic criteria, we recognize some inconsistencies in the divergence thresholds used. Phylopart performs clustering only at the tip level, but DYNAMITE overcomes this problem and better resembles the dynamic lineage/clade assignment. Both Phylopart and DYNAMITE can be biased by sampling rates [26].

All SARS-CoV-2 phylogenies analyzed in this work show high tree imbalance, suggesting short infection times and antigenic drift [27], perhaps due to progressive transition from innate to adaptive immunity in the population [28]. Since there was not a single tree that had the highest imbalance across all measures, we cannot determine if one out of the global, regional or clade-specific epidemics might be subject to higher antigenic drift than the others.

Limitations of this work include assumptions' violations of a correct phylogeny and of a uniform, representative sampling [29]. A possible solution to overcome sampling bias could be to generate multiple phylogenies using the TARDiS method [30], which optimizes both genetic diversity and temporal distribution, and perform ensemble analyses. Another problem is that we assumed 100% node reliability for the Nextstrain's trees, whereas poorly supported subtrees might have impacted the cluster agreement calculation.

In conclusion, we foresee improvements in the integration of statistical criteria based on phylogenetic diversity for lineage determination as well as identification of clusters of concerns within sub-epidemics, bearing always the necessity of representative sampling.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

This work is made possible by the open sharing of genetic data by research groups from all over the world. We gratefully acknowledge their contributions (see online supplement for full credits).

This work was supported in part by NSF grant #2028221 and NIH grant # R01AI170187.

References

[1]. Harvey WT et al. , "SARS-CoV-2 variants, spike mutations and immune escape," Nat Rev Microbiol, vol. 19, no. 7, Art. no. 7, Jul. 2021, doi: 10.1038/s41579-021-00573-0. [PubMed: 33219332]

[2]. Rochman ND, Wolf YI, Faure G, Mutz P, Zhang F, and Koonin EV, "Ongoing global and regional adaptive evolution of SARS-CoV-2," Proceedings of the National Academy of Sciences, vol. 118, no. 29, p. e2104241118, Jul. 2021, doi: 10.1073/pnas.2104241118.

[3]. Campbell F et al. , "Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021," Euro Surveill, vol. 26, no. 24, p. 2100509, Jun. 2021, doi: 10.2807/1560-7917.ES.2021.26.24.2100509. [PubMed: 34142653]

[4]. Khan K et al. , "Omicron BA.4/BA.5 escape neutralizing immunity elicited by BA.1 infection," Nat Commun, vol. 13, no. 1, p. 4686, Aug. 2022, doi: 10.1038/s41467-022-32396-9. [PubMed: 35948557]

[5]. Martin DP et al. , "Selection analysis identifies unusual clustered mutational changes in Omicron lineage BA.1 that likely impact Spike function," bioRxiv, p. 2022.01.14.476382, Jan. 2022, doi: 10.1101/2022.01.14.476382.

[6]. Gutierrez B et al. , "Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America," Cell Host Microbe, vol. 30, no. 8, pp. 1112–1123.e3, Aug. 2022, doi: 10.1016/j.chom.2022.06.010. [PubMed: 35853454]

[7]. Rambaut A et al. , "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology," Nat Microbiol, vol. 5, no. 11, Art. no. 11, Nov. 2020, doi: 10.1038/s41564-020-0770-5.

[8]. Subissi L et al. , "An early warning system for emerging SARS-CoV-2 variants," Nat Med, vol. 28, no. 6, Art. no. 6, Jun. 2022, doi: 10.1038/s41591-022-01836-w. [PubMed: 34992264]

[9]. Hadfield J et al. , "Nextstrain: real-time tracking of pathogen evolution," Bioinformatics, vol. 34, no. 23, pp. 4121–4123, Dec. 2018, doi: 10.1093/bioinformatics/bty407. [PubMed: 29790939]

[10]. Dellicour S et al. , "A Phylodynamic Workflow to Rapidly Gain Insights into the Dispersal History and Dynamics of SARS-CoV-2 Lineages," Mol Biol Evol, vol. 38, no. 4, pp. 1608–1613, Apr. 2021, doi: 10.1093/molbev/msaa284. [PubMed: 33316043]

[11]. Puccio B, Lomoio U, Paola LD, Guzzi PH, and Veltri P, "Annotating Protein Structures for Understanding SARS-CoV-2 Interactome," p. 10.

[12]. Attwood SW, Hill SC, Aanensen DM, Connor TR, and Pybus OG, "Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic," Nat Rev Genet, vol. 23, no. 9, Art. no. 9, Sep. 2022, doi: 10.1038/s41576-022-00483-8.

[13]. Kumar Das J, Tradigo G, Veltri P, H Guzzi P, and Roy S, "Data science in unveiling COVID-19 pathogenesis and diagnosis: evolutionary origin to drug repurposing," Briefings in Bioinformatics, vol. 22, no. 2, pp. 855–872, Mar. 2021, doi: 10.1093/bib/bbaa420. [PubMed: 33592108]

[14]. Alkhamis MA, Fountain-Jones NM, Khajah MM, Alghounaim M, and Al-Sabah SK, "Comparative phylodynamics reveals the evolutionary history of SARS-CoV-2 emerging variants in the Arabian Peninsula," Virus Evol, vol. 8, no. 1, p. veac040, May 2022, doi: 10.1093/ve/veac040. [PubMed: 35677574]

[15]. Serwin K, Aksak-Wąs B, and Parczewski M, "Phylodynamic Dispersal of SARS-CoV-2 Lineages Circulating across Polish–German Border Provinces," Viruses, vol. 14, no. 5, p. 884, Apr. 2022, doi: 10.3390/v14050884. [PubMed: 35632625]

[16]. Lemieux JE et al. , "Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events," Science, vol. 371, no. 6529, p. eabe3261, Feb. 2021, doi: 10.1126/science.abe3261. [PubMed: 33303686]

[17]. Giovanetti M et al. , "Genomic epidemiology of the SARS-CoV-2 epidemic in Brazil," Nat Microbiol, vol. 7, no. 9, pp. 1490–1500, Sep. 2022, doi: 10.1038/s41564-022-01191-z. [PubMed: 35982313]

[18]. McCrone JT et al. , "Context-specific emergence and growth of the SARS-CoV-2 Delta variant," Nature, Aug. 2022, doi: 10.1038/s41586-022-05200-3.

[19]. Prosperi MCF et al. , "A novel methodology for large-scale phylogeny partition," Nat Commun, vol. 2, no. 1, Art. no. 1, May 2011, doi: 10.1038/ncomms1325.

[20]. Magalis BR, Marini S, Salemi M, and Prosperi M, "DYNAMITE: a phylogenetic tool for identification of dynamic transmission epicenters." bioRxiv, p. 2021.01.21.427647, Jan. 22, 2021. doi: 10.1101/2021.01.21.427647.

[21]. Rand WM, "Objective Criteria for the Evaluation of Clustering Methods," Journal of the American Statistical Association, vol. 66, no. 336, pp. 846–850, Dec. 1971, doi: 10.1080/01621459.1971.10482356.

[22]. Fischer M, Herbst L, Kersting S, Kühn L, and Wicke K, "Tree balance indices: a comprehensive survey." arXiv, Sep. 25, 2021. doi: 10.48550/arXiv.2109.12281.

[23]. Norström MM, Prosperi MCF, Gray RR, Karlsson AC, and Salemi M, "PhyloTempo: A Set of R Scripts for Assessing and Visualizing Temporal Clustering in Genealogies Inferred from Serially Sampled Viral Sequences," Evol Bioinform Online, vol. 8, p. EBO.S9738, Jan. 2012, doi: 10.4137/EBO.S9738.

[24]. Poon AFY, Walker LW, Murray H, McCloskey RM, Harrigan PR, and Liang RH, "Mapping the Shapes of Phylogenetic Trees from Human and Zoonotic RNA Viruses," PLOS ONE, vol. 8, no. 11, p. e78122, Nov. 2013, doi: 10.1371/journal.pone.0078122. [PubMed: 24223766]

[25]. Colijn C and Gardy J, "Phylogenetic tree shapes resolve disease transmission patterns," Evol Med Public Health, vol. 2014, no. 1, pp. 96–108, Jun. 2014, doi: 10.1093/emph/eou018. [PubMed: 24916411]

[26]. Poon AFY, "Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks," Virus Evolution, vol. 2, no. 2, p. vew031, Jul. 2016, doi: 10.1093/ve/vew031. [PubMed: 28058111]

[27]. Grenfell BT et al. , "Unifying the Epidemiological and Evolutionary Dynamics of Pathogens," Science, vol. 303, no. 5656, pp. 327–332, Jan. 2004, doi: 10.1126/science.1090727. [PubMed: 14726583]

[28]. "Antigenic drift: Understanding COVID-19 - PMC." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8669911/ (accessed Sep. 15, 2022).

[29]. Stack JC, Welch JD, Ferrari MJ, Shapiro BU, and Grenfell BT, "Protocols for sampling viral sequences to study epidemic dynamics," J R Soc Interface, vol. 7, no. 48, pp. 1119–1127, Jul. 2010, doi: 10.1098/rsif.2009.0530. [PubMed: 20147314]

[30]. Marini S, Mavian C, Riva A, Prosperi M, Salemi M, and Rife Magalis B, "Optimizing viral genome subsampling by genetic diversity and temporal distribution (TARDiS) for phylogenetics," Bioinformatics, vol. 38, no. 3, pp. 856–860, Feb. 2022, doi: 10.1093/bioinformatics/btab725. [PubMed: 34672334]

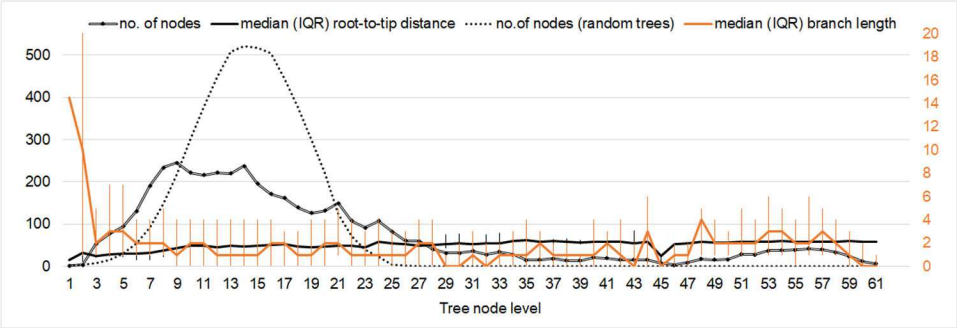

Fig. 1.

Distribution of number of nodes compared to random trees (left Y axis), median (IQR) root-to-tip distance (right Y axis), and median (IQR) branch length (secondary axis) per tree level (X axis) for the SARS-CoV-2 global phylogeny.

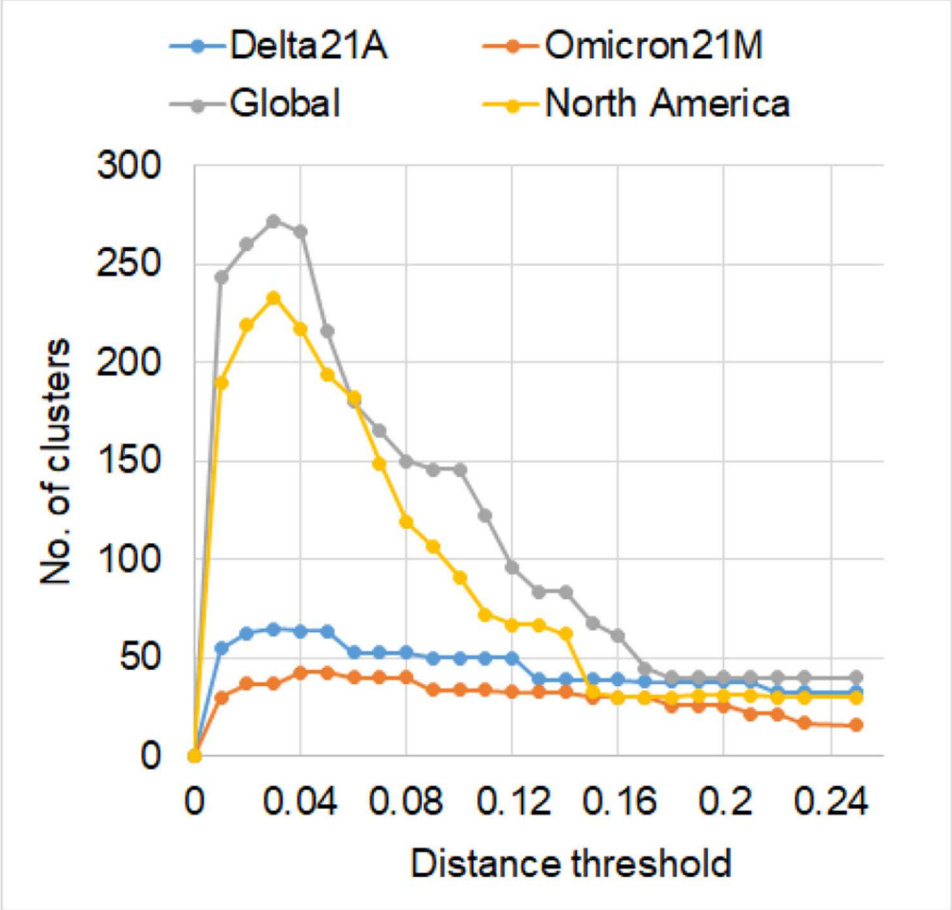

Fig. 2.

Number of transmission clusters found by Phylopart on the global, North American, Delta21A, and Omicron20M SARS-CoV-2 phylogenies by varying the patristic distance threshold.

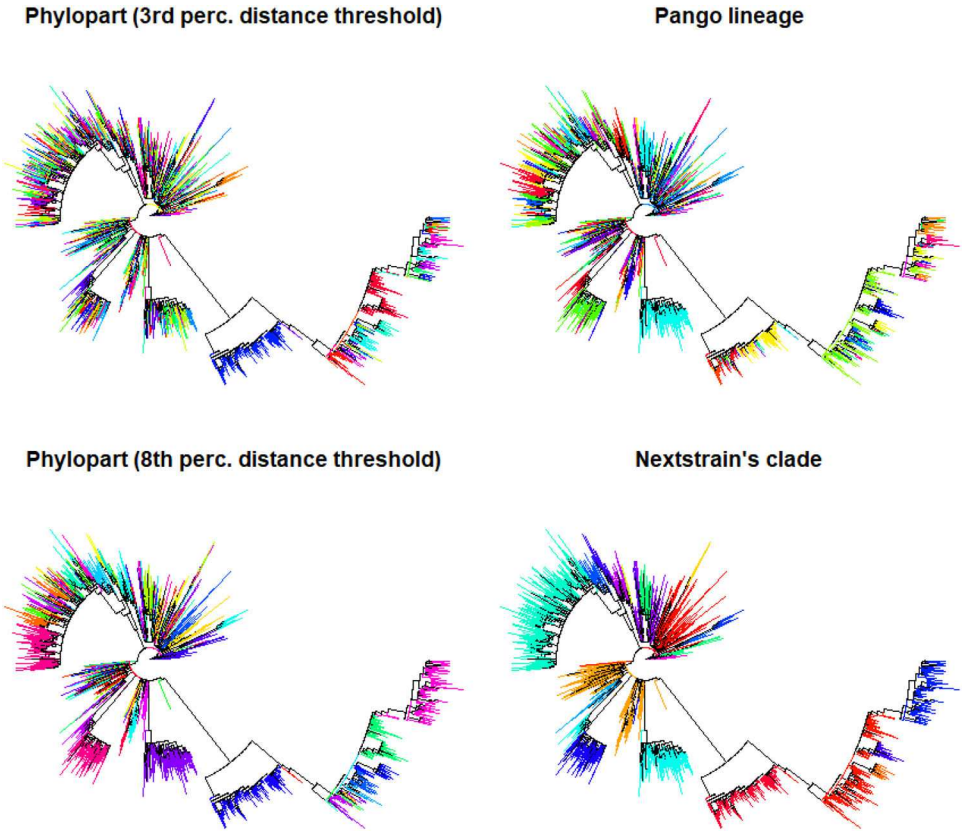

Fig. 3.

Comparison of phylogenetic clustering by Phylopart with Pango lineages and Nextstrain's clades on the SARS-CoV-2 global tree.