Phylogenetic Analysis to Detect COVID Superspreaders

ABSTRACT

Aims: Detection of superspreading events by phylogenetic analysis of nucleotide sequences from a population of individuals collected from a narrow time interval.
Study design: Retrieve nucleic acid sequences, construct multiple sequence alignments, and build phylogenetic networks to determine sources of infection.
Place and Duration of Study: Sample: From GIS AID. Place of Analysis: Delaware Biotechnology Institute. Period of study: June-August, 2022.
Methodology: Sequences for analysis were sampled from the GISAID initiative's open access SARS-CoV-2 genome database. We selectied for high quality nucleotide sequences submitted by Delaware labs between March 18 and April 14, 2021, an important time span of 4 weeks which saw the Alpha variant spread rapidly in the Delaware population.
Results: Four sources accounted for 155 of the 401 sequences. In other words, 39% of all cases were rooted in just four sources.
Conclusion: Thus, superspreading seems to have a major impact upon the proportion of individuals in a population affected with COVID.

Comment [S1]: Kindly check the spelling.

Comment [S2]: Kindly explain this is 155 or 166 or 186?

Keywords: COVID, superspreaders, phylogenetic networks,

1. INTRODUCTION

> "It is now generally thought that superspreading is very common in epidemics, with a rough rule of thumb being that 20% of a population causes 80% of disease cases." (Brauer, 2019 abc) .

Contact tracing by interviewing patients infected with a virus has long been a critical aspect of public health approaches to epidemiology. Since the development of sequencing of pathogen genomes and of powerful mathematical and computational procedures for inferring the evolutionary history of the spread of infections, we have a more direct method of inferring who was infected by whom. Popa *et al.* (2020) noted that "Superspreading events shaped the coronavirus disease 2019 (COVID-19) pandemic." They reported that: "Our results integrating epidemiological and sequencing data emphasize that phylogenetic analyses of SARS-CoV-2 sequences empower robust tracing from interindividual to local and international spreading events. … This study underscores the value of combining epidemiological approaches with virus genome sequencing to provide critical information to help public health experts track pathogen spread." While a follow-up study by Martinand Koelle(2021) was critical of some of Popa ***et al.'s*** interpretations, they concluded that: "Small bottleneck sizes also mean that infections generally start off with very little, if any, viral genetic diversity, such that acute infections will likely be characterized by low levels of viral diversity except in instances of superinfectionconsistent with other recent studies."We believe that the results of these two studies and others (Edholm*et al.*, 2018) make it both

easier and highly beneficial to examine other local populations to determine the impact of superspreading more generally. Therefore, we examined a sample from our state of Delaware because we felt that three important criteria are met: (a) a sufficient set of sequence data had been collected to have a reasonable size; (b) the sequences were available for over a period of rapid spread of the disease; and (c) a new variant occurred which was rapidly spreading during a short time frame.

Unfortunately few epidemiological studies account for the significant role of superspreading. In particular,phylogenetic detection of superspreading is understudied particularly when insufficient sequencing is monitoring the course of infection in populations. Only by collecting and evolutionarily analyzing the sequences from the viruses can we infer the fine-scale dynamics of viral spread.

METHODOLOGY:

Phylogenetic analysis is able to contribute to epidemiological studies in six major different ways (Jungck *et al.*, 2006): (1) determining the origin of a pandemic; (2) identifying new variants as containing sufficiently different mutations such their they have different levels of infectivity, morbidity, and mortality; (3) determining when such variants evolved (Wang *et al*, 2020); (4) determining the rate of mutation (Chakrabort*yet al.*, 2021 (Figure 1) and Robeva and Jungck, 2023); (5) the intensity of selection; and (6) determining where such variants evolved. These investigations fall into three categories identified in different literatures with different taxonomic names that focus on: (1) time: molecular clocks or chronocladistics; (2) space: phylogeography or topocladistics; and, (3) genealogy: common ancestry or patrocladistics. Associated with each inquiry are a variety of different phylogenetic methods. For example, in order to build a molecular clock, an important assumption about the distance matrix of differences between sequences inferred from a multiple sequence alignment should satisfy an ultrametric condition; namely, that. The rate of mutation is constant over time. If this assumption is satisfied, many tree builders use the UPGMA (Unweighted Pair Group Method with Arithmetic Mean)algorithm. If the distance matrix doesn't satisfy this strict assumption, often the Neighbor Joining algorithm is employed to infer the genealogy. MEGA (Molecular Evolutionary Genetic Analysis <https://www.megasoftware.net/>) is one of the most widely used software for conducting such analyses as well as using a Maximum Likelihood assessment of whether particular bifurcations in the tree are well supported well or not. If one wants to look at specific events (such as transitions, transversions, deletions, inversions, etc.) have occurred, a character based or cladistic approach. The software Mesquite (<https://www.mesquiteproject.org/>) is appropriate for such analyses. When many assumptions about the distance matrix are not met, it is often important to ask how tree-ed is your data. Thus, instead of forcing the data to fit a tree (which is a planar graph), we can keep the distances in a multidimensional phylogenetic network. Phylogenetic networks are better than phylogenetic trees when particularly important biological features thought to underlie viral evolution such as recombination and horizontal gene transfer occur.Because several assumptions were not met in our distance matrices from our multiple sequence alignments, we built a phylogenetic network using the Splitstree(https://uni-tuebingen.de/en/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/splitstree/) software (Huson and Bryant, 2006).

Comment [S3]: Why underline?

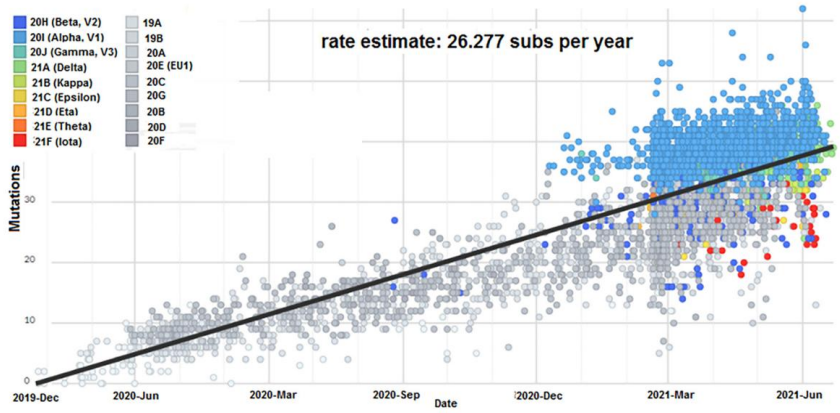

Figure 1. The mutation rate of the COVID 19 virus was determined by Chakrabortyet *al.* (2021) to be over 26 substitutions per genome per year. Their legend: "Scatterplot showing the genome diversity cluster of all circulating lineages between December 2019 and June 2021 .. through the Nextstrain server, using GISAID data." Creative Commons license level 4.

Choice of period of study: Early on we had access to the work of Hockstein *et al.* (2021) who examined the spread of infection in Delaware. They were able to show that students in a historically black university were safer on campus than in the general population before vaccines became available by doing masking, social distancing, contact tracing, and frequent testing. In their comparison with campus and state data, a second peak of the pandemic occurred in Delaware and on their campus between March 18 and April 14, 2021 (Figure 2).

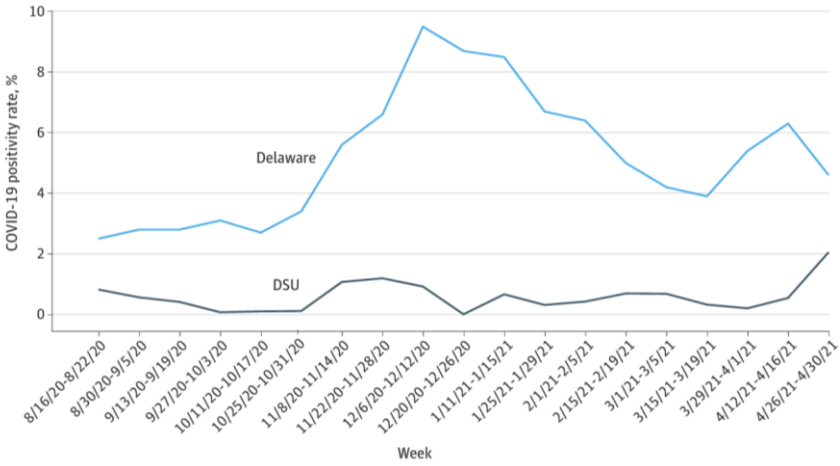

Figure 2. COVID-19 positivity rates of Delaware State University population versus that in the general Delaware population (Hockstein *et al.*, 2021; personal permission).

Source of sequences: GISAID provides open access to sequence data on the coronavirus causing COVID-19 from around the world (<https://gisaid.org/>). We selected a genetically diverse dataset of 401 nucleotide sequences. The full genome is roughly 30 kilobases. We decided to focus on the spike protein as it is the primary target of vaccines (Zhu *et al.*, 2021). Thus we chose a region of 3.8k bases which covered the spike protein gene. The sequences were aligned with the MUSCLE (<bi.ac.uk/Tools/msa/muscle/>) multiple sequence alignment tool and the resulting data matrix was entered into the SPLITSTREE software to generate a phylogenetic network.

RESULTS:

Our phylogenetic network of 401 sequences from Delaware COVID patients over the period betweenMarch 18 and April 14, 2021collected from the GISAID database of the region of the COVID virus genome covering the spike protein is shown in Figure 3.

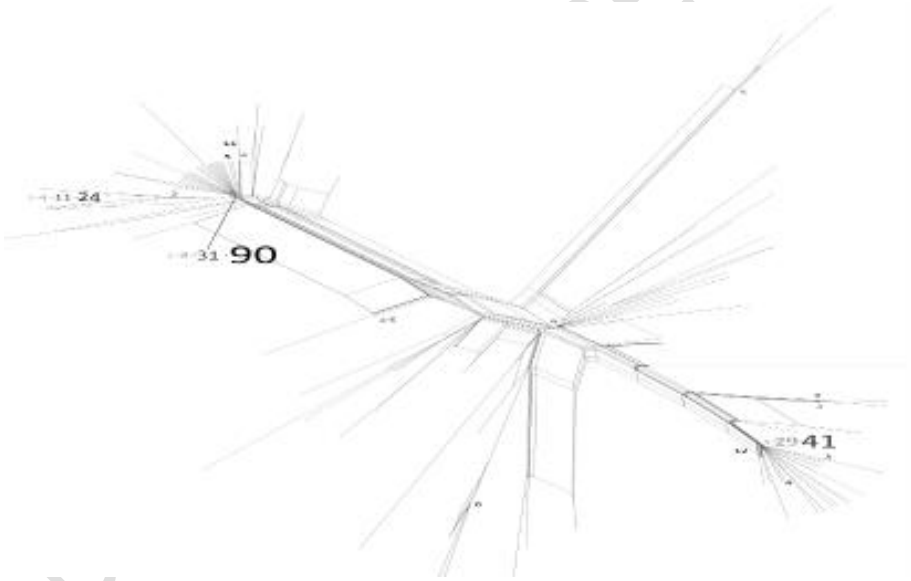

Figure 3. Splits tree generated phylogenetic network of 401 sequences from Delaware COVID patients over the period March 18 to April 14, 2021collected from the GISAID database.

If we take the four major branches of duplicate sequences (90 + 41 + + 31 + 24), these account for 166 out of the 401 cases, or 41% of the cases. This is far above Brauer's (2019 abc) estimate of 20%. We believe that this is consistent with the hypothesis of superspreaders being responsible for a significant fraction of infections in a restricted population in a short interval of time.

Comment [S4]: Kindly explain.

Comment [S5]: Kindly explain is it 155 or 166 or 186?

Conclusion:

Within a genetically diverse dataset of 401 nucleotide sequences, large amounts of duplicate Spike sequences were present. Since we were investigating fairlylong lengths—3.8kilobases—of these nucleotide sequences, we believe that the presence of somany identical sequences is consistent with the hypothesis that they originated from the same or highly similar genetic sources and hence are evidence of superspreading.

Super-spreading provides a plausible explanation of the large number of identical sequences observed in the analysis, positing, for example, that many of the 90 infectious cases highlighted in week 4 of Figure3 could be traced back to an earlier case which may have been one recorded in a previous week.If this data were corroborated by patient data and contact tracing of identical sequences, it would confirm the likely hypothesis that super-spreading is visible at the nucleotide level of a dataset and can be identified using phylogenetic analysis.Franke*et al.* (2022) subsequent to our work give a larger framework for the transmission of COVID-19 in Delaware but they do not explicitly address the issue of superspreaders.

These findings provide grounds for investment in future studies assessing whether phylogenetic analysis could be used in the estimation of contact tracing based off sequence data rather than patient data.

Subsequent to our work,Taube, Miller, and Drake (2022) published "An open-access database of infectious disease transmission trees to explore superspreader epidemiology." Their work will make it much easier for individuals to build upon our and others' work in phylogenetic analysis of COVID transmission. Such studies are crucial for setting public health policy.

In addition, we are aware that there is some controversy about the use of phylogenetic network analysis. Chookajornraised an alarm about a previous phylogenetic network analysis by Foster *et al.* (2020a):
"Asanevolutionarybiologistworkinginadevelopingcountry,Ihaveexperiencedfirsthandhowsensationalfindingscaninfluencedecision-makingprocesses bydivertingtimeandresourcestocontrolvirusstrains deemedtobe"moreaggressive."Inthefogofwar, scarceresourcesareallocatedinhaste,andthedevelopingworlddoesnothavewell-informedscienceadvisers sittingineverykeymeetingtohelpprovidebalanced scientificviewpoints.Thescientificcommunity,asa whole,needstobeextracautiousininterpretingnew findingsrelatedtocoronavirusdisease2019(COVID19),andany potential misinformationmustbepromptly addressed." While Foster *et al.* (2020b) submitted a rejoinder, we believe that the sensitivities such important policy ramifications need heterogeneous stakeholders with multiple perspectives to be at the decision table.

Comment [S6]: Kindly justify proper format.

Finally, we are particularly worried that since public concern about the ongoing COVID pandemic has decreased and concomitantly there has been less funding for sequencing current strains in the broad international community that we are facing a situation where even doing phylogenetic analysis of available sequences may be insufficient to identify many newly evolving variants and their associated levels of infectivity, morbidity, and mortality.

References:

Attwood, S.W., Hill, S.C., Aanensen, D.M. *et al.* (2022).“Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic.”*Nat Rev Genet.* **23**, 547–562 (2022). https://doi.org/10.1038/s41576-022-00483-8

Brauer, Fred. (2019). “Viral Math: For hundreds of years, mathematical epidemiology has helped us understand how diseases spread and what treatments will be effective against them.” *Tablet Magazine* https://www.tabletmag.com/sections/science/articles/viral-epidemiology

Brauer, Fred. (2019). "The final size of a serious epidemic." *Bulletin of Mathematical biology* 81 (3): 869-877.

Brauer, Fred.(2019). Early estimates of epidemic final sizes. *Journal of Biological Dynamics,* 13(sup1), 23-30.

Chakraborty, Chiranjib, Ashish Ranjan Sharma, Manojit Bhattacharya, Govindasamy Agoramoorthyand, Sang-Soo Lee.(2021). Evolution, Mode of Transmission, and Mutational Landscape of Newly Emerging SARS-CoV-2 Variants. *mBio*12 (*4*): e01140-21 (22 pages).

Galvani, Alison P., and Robert M. May. (2005). "Dimensions of superspreading." *Nature* 438 (7066): 293-295.

Edholm, Christina J., Blessing O. Emerenini, Anarina L. Murillo, Omar Saucedo, Nika Shakiba, Xueying Wang, Linda JS Allen, and Angela Peace. (2018)."Searching for superspreaders: Identifying epidemic patterns associated with superspreading events in stochastic models." *Understanding complex biological systems with mathematic*s (2018): 1-29.

Forster, Peter, Lucy Forster, Colin Renfrew, and Michael Forster. PhylogeneticnetworkanalysisofSARS-CoV-2genomes*. Proc.Natl.Acad.Sci.U.S.A.* 117,9241–9243(2020a).

Forster, Peter, Lucy Forster, Colin Renfrew, and Michael Forster. (May 21, 2020b). “Reply to Sánchez-Pacheco et al., Chookajorn, and Mavian et al.: Explaining phylogenetic network analysis of SARS-CoV-2 genomes.” *Proceedings of the National Academy of Sciences* 117 (23) 12524-12525.

Franke, K. R., Isett, R., Robbins, A., Paquette-Straub, C., Shapiro, C. A., Lee, M. M., &Crowgey, E. L. (2022). Genomic surveillance of SARS-COV-2 in the state of Delaware reveals tremendous genomic diversity. PLOS ONE, 17(1).

Galvani, Alison P., and Robert M. May. "Dimensions of superspreading." *Nature* 438, no. 7066 (2005): 293-295.

Goyal, Ashish, Daniel Reeves, and Joshua T. Schiffer. "Early super-spreader events are a likely determinant of novel SARS-CoV-2 variant predominance." medRxiv (2021): 2021-03.

Hasan, Agus, Hadi Susanto, Muhammad Firmansyah Kasim, Nuning Nuraini, Bony Lestari, DessyTriany, and WidyastutiWidyastuti. "Superspreading in early transmissions of COVID-19 in Indonesia." Scientific reports 10, no. 1 (2020): 1-4.

Hockstein, Neil G., LaKresha Moultrie, Michelle Fisher, R. Christopher Mason, Derrick C. Scott, Joan F. Coker, Autumn Tuxward*et al*. "Assessment of a multifaceted approach, including frequent PCR testing, to mitigation of COVID-19 transmission at a residential historically Black university." *JAMA Network Open* 4, no. 12 (2021): e2137189-e2137189.

Huson, Daniel H., and D. Bryant. (2006). Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution, 23(2), 254–267. https://doi.org/10.1093/molbev/msj030

James, Alex, Jonathan W. Pitchford, and Michael J. Plank. "An event-based model of superspreading in epidemics." Proceedings of the Royal Society B: Biological Sciences 274, no. 1610 (2007): 741-747.

Jha, S., Kumar, S., & Rai, S. K. (2020). Significance of super spreader events in covid-19. Indian Journal of Public Health, 64(6), 139.

Jungck, J. R., N. Khiripet, R. Viruchpinta, and J. Maneewattanapluk. (2006): "Evolutionary bioinformatics: making meaning of microbes, molecules, maps." *MICROBE Magazine (American Society for Microbiology* 1 (8): 365-371.

Khare, S., Gurry, C., Freitas, L., B Schultz, M., Bach, G., Diallo, A., Akite, N., Ho, J., TC Lee, R., Yeo, W., Core Curation Team, G. I. S. A. I. D., & Maurer-Stroh, S. (2021). Gisaid's role in pandemic response. China CDC Weekly, 3(49), 1049–1051.

Li, Xiao-Ping, Saif Ullah, Hina Zahir, Ahmed Alshehri, Muhammad Bilal Riaz, and Basem Al Alwan. "Modeling the dynamics of coronavirus with super-spreader class: A fractal-fractional approach." Results in Physics 34 (2022): 105179.

Lloyd-Smith, James O., Sebastian J. Schreiber, P. Ekkehard Kopp, and Wayne M. Getz. "Superspreading and the effect of individual variation on disease emergence." *Nature* 438, no. 7066 (2005): 355-359.

Martin, Michael A., and Katia Koelle. "Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2"." *Science translational medicine* 13, no. 617 (2021): eabh1803.

McCaig, Chris, Mike Begon, Rachel Norman, and Carron Shankland. "A symbolic investigation of superspreaders." *Bulletin of Mathematical Biology* 73, no. 4 (2011): 777-794.

Müller, Johannes, and Volker Hösel. "Contact tracing & super-spreaders in the branching-process model." *Journal of Mathematical Biology* 86, no. 2 (2023): 24.

Ndaïrou, F., Area, I., Nieto, J. J., & Torres, D. F. M. (2020). Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. Chaos, Solitons & Fractals, 135, 109846.

Popa, Alexandra, Jakob-Wendelin Genger, Michael D. Nicholson, Thomas Penz, Daniela Schmid, Stephan W. Aberle, Benedikt Agerer et al. "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2." *Science translational medicine* 12, no. 573 (2020): eabe2555.

Robeva, Raina S., and John R. Jungck. "Fascination with Fluctuation: Luria and Delbrück's Legacy." *Axioms* 12, no. 3 (2023): 280.

Sánchez-Pacheco, Santiago J., Sungsik Kong, Paola Pulido-Santacruz, and Laura Kubatko. (May 7, 2020). "Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary." *Proceedings of the National Academy of Sciences* 117 (23) 12518-12519.

Santana-Cibrian, Mario, Manuel A. Acuna-Zegarra, and Jorge X. Velasco-Hernandez. "Lifting mobility restrictions and the effect of superspreading events on the short-term dynamics of COVID-19." *Mathematical Biosciences and Engineering* 17, no. 5 (2020): 6240-6258.

Tamura, K., Stecher, G., & Kumar, S. (2021). Mega11: Molecular Evolutionary Genetics Analysis Version 11. Molecular Biology and Evolution, 38(7), 3022–3027.

Taube, Juliana C., Paige B. Miller, and John M. Drake. "An open-access database of infectious disease transmission trees to explore superspreader epidemiology." *PLoS Biology* 20, no. 6 (2022): e3001685.

Wang, Jann-Tay, You-Yu Lin, Sui-Yuan Chang, Shiou-Hwei Yeh, Bor-Hsian Hu, Pei-Jer Chen, and, Shan-Chwen Chang. (2020). "The role of phylogenetic analysis in clarifying the infection source of a COVID-19 patient." *Journal of Infection* 81 (1): 147-178.

Zenk, Lukas, Gerald Steiner, Miguel Pina e Cunha, Manfred D. Laubichler, Martin Bertau, Martin J. Kainz, Carlo Jäger, and Eva S. Schernhammer. "Fast response to superspreading: uncertainty and complexity in the context of COVID-19." *International journal of environmental research and public health* 17, no. 21 (2020): 7884.

Zhu C, He G, Yin Q, Zeng L, Ye X, Shi Y, Xu W. Molecular biology of the SARs-CoV-2 spike protein: A review of current knowledge. *J Med Virol.*(2021) 93(10):5729-5741. doi: 10.1002/jmv.27132.