Defining and analyzing differences in state-by-state and year-by-year SARS-CoV-2 transmission clusters up to May 10, 2023

Benjamin Cai

Institute for Computing in Research

July 2024

Abstract

Identifying transmission clusters is integral to limiting viral spread. SARS-CoV-2 is unique in its quick spread and the resultant number of sequences collected given the time frame of its existence. By analyzing the transmission clusters of SARS-CoV-2 on a state-by-state basis and determining accurate limits in what is a transmission clusters, we can compare differences in state size and policy on SARS-CoV-2 spread. The data collected can then be extrapolated to other viruses similar in transmission to SARS-CoV-2.

***Keywords**: SARS-CoV-2, COVID-19, transmission networks, transmission clusters, phylogenetics*

1 Introduction

Understanding the movement of viruses has been fundamental to combating them. Within the past 5 years, SARS-CoV-2 has infected an estimated 700 million across the world and an estimated 100 million within the U.S., including 7 million and 1 million deaths, respectively [1]. SARS-CoV-2 is extremely transmissable with the primary route of transmission being respiratory droplets from infected persons [2]. Currently, it is widely accepted that SARS-CoV-2 is spread through people in close contact such as family members or friends at a conversational distance [2].

Transmission clusters typically represent groups of infected individuals [3]. These individuals are then represented on a phylogenetic tree as a single sequence positioned at a tree tip. These tree tips hold corresponding metadata such as location, date, etc. By finding the MCRA (most common recent ancestor) of sequences (tips) matching by location, we can identify clades within the phylogeny that correspond to geographic transmission clusters.

Due to SARS-CoV-2 worldwide impact, sequencing efforts and sequencing technology exploded [4]. Fortunately, much of this data can be curated and accessed freely on GenBank. However, the problem that comes into play is computing power. Due to the exponential requirements of tree-building, it is oftentimes better to sample large portions of sequences down into smaller subsamples. In this paper, the problem of subsampling producing inadequate representations of the data is remedied by the size of the subsample in relation to the number of geographic divisions themselves. Even in the smallest subsample, there are still over fifty strains for each state.

Once the tree is created, it can be loaded into phylogenetic analysis software such as the NextStrain API, ape, ITOL, etc. In this case, we have created an open source tool to automatically build transmission clusters on top of an existing tree. Unfortunately, this process is exponential depending on the total number of nodes and tips.

2 Tree-building algorithms and Identification of Clusters

Genetic distance and the tree itself was automatically calculated by the NextStrain API [5]. After the tree was created, the tree (newick) and its corresponding metadata was plugged into R. Using the ape library, a list of subtrees was computed. Next, the total percent and number of tips corresponding to the geographic region in question for each individual subtree was calculated. This result which was called percentExcluded and numberExcluded was used to determine definitions for transmission clusters. Then, the size of those clusters would be linked to dates by the libraries castor and phytools.

3 Dataset

Oftentimes, the most limiting aspect within phylogenetic analysis is acquiring the data itself. The two datasets used here were found firstly by taking the Full Open metadata.tsv file from the NextStrain remote inputs download. Next, that metadata file was filtered by state and date to match up with the number of cases per year per state by a ratio of 0.0001 : 1. Next, the genbank accession numbers of those new filtered metadata files were read into a .txt file. Then, that .txt file was split by batches of 10,000 accession numbers and read into the batch entrez support offered by GenBank to get .fasta files. The .fasta files were concatenated and the names for each sequence was renamed based on the sequence metadata file. The cutoffs for each year were 01-16-2020, 05-13-2020, 05-12-2021, 05-11-2022, and 05-10-2023.

The dataset corresponds directly to number of cases per state per year by a ratio of 0.0001 : 1. This means if one state, say California had much more cases in 2022, then the corresponding number of sequences between 05-11-2022 and 05-12-2021 would be much higher than in other years and other states with less cases. As a direct example from the data, Alaska between 01-16-2020 and 05-13-2020 had 389 cases of SARS-CoV-2, California between 05-12-2021 and 05-11-2022 had 5,489,382 cases, and New Mexico between 05-11-2022 and 05-10-2023 had 155,994 cases (Figure 1 & 3). Thus, in this example, Alaska would be represented by 1 case, California by 549, and New Mexico by 16 (Figure 2 & 4). The purpose of this dataset is to determine how much cases correlate to transmission clusters, and it has size 9,889.

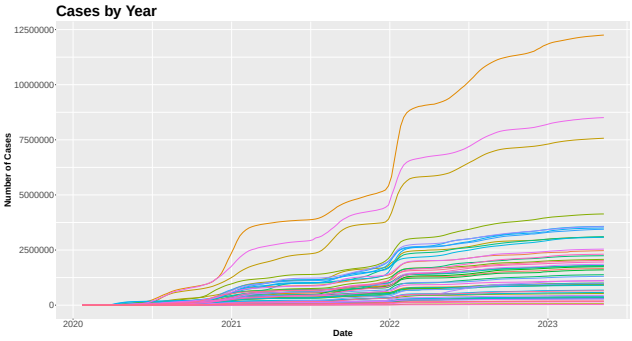

Figure 1: *Cumulative Cases*

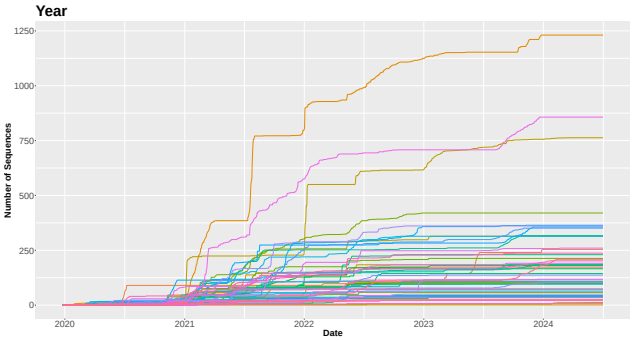

Figure 2: *Cumulative Sequences*

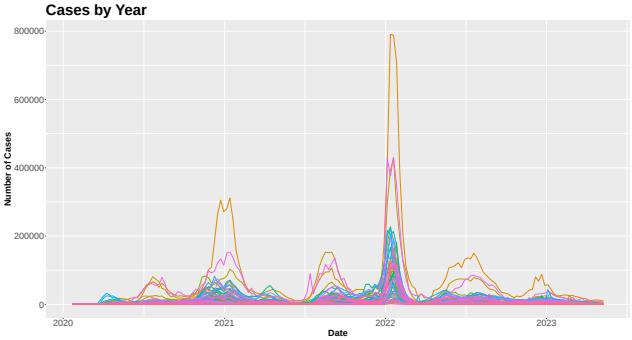

Figure 3: *New Cases*

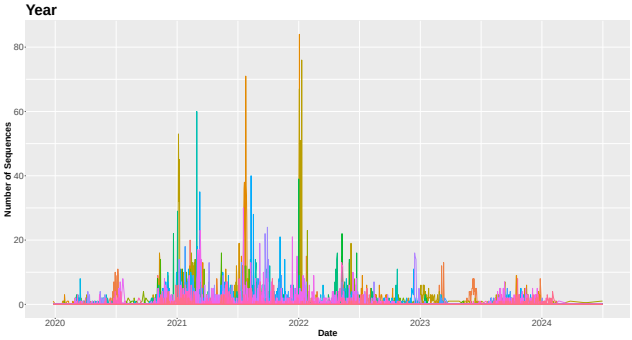

Figure 4: *New Sequences*

Figure 4 does not exactly match up with figure 3 due to subsampling that focused on grouping sequences by year rather than by each month or week. This departure from following the cases absolutely was due to a difficulty in acquiring sequences during the peak weeks. However, even then, Figures 1 and 2 do match up extremely closely.

One option often used in phylogenetic analyses of these kinds is to insert in contextual sequences [5]. For example, for this study, a contextual sequence group would be 800 sequences that are genetically similar to sequences within dataset 1. The purpose of these contextual sequences is to link up sequences that may become genetically closer with another sequence. In this study, it was decided to omit global contextual sequences as to focus solely on transmission clusters within the U.S., rather than from globally into the U.S.

4 Results

For each state of the U.S., an individual analysis was done on each subtree. Then, these subtrees were compared to see the highest percentage of tips from that state. In these comparisons, each subtree compared would need to have at least 5 tips total and 50% of all tips being from the state of interest in order to be counted as a transmission cluster. Without the size distinction, a single pair of tips from the same state could easily overrun the model due to those two tips being 100% out of a subtree of size two. The 50% distinction exists in order to filter out subtrees that are simply too diluted to be considered true transmission clusters. In the tree themselves, tips with sequences from the state of interest would be highlighted in pink while the transmission clusters themselves would be

highlighted by a red dot whose brightness correlates to the percent of the transmission cluster. Lastly, a blue box was added over the largest transmission cluster found. The largest transmission clusters would be ranked firstly by percent, then by size in the event of ties.

Most of the state trees found can be divided into 4 major categories. The first are states that have too little samples to find a true transmission cluster (Figure 5). These are the states who have too little cases in relation to other states to the point that there are not enough samples concentrated to see any true result. Alaska in Figure 5 only had 31 samples in total, and thus, the biggest transmission cluster would not even be considered a transmission cluster according to our 50% and 5 occurrence minimum. The second category are states that have enough samples to see transmission clusters but not enough samples to see mega transmission clusters (Figure 6). It is these two first categories that make up the majority of states. The third category are states that have the highest number of samples, and thus, have a high amount and size of transmission clusters (Figure 7). California had around 1200 samples, which was the maximum number out of any state. The final category are reserved for states with mega clusters. A mega cluster can be categorized as a transmission cluster with >80% and >50 occurrences (Figure 8). Surprisingly, Texas did not have the most samples sitting at only 800, but still had the largest transmission cluster by far.

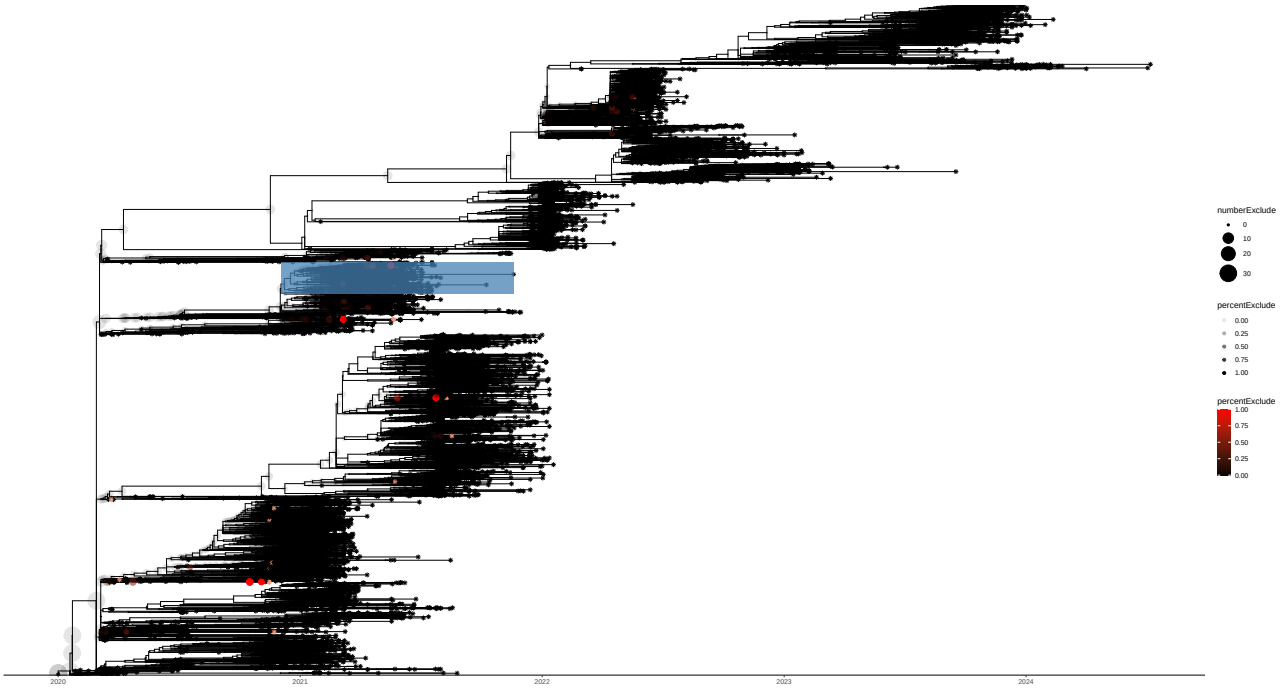

Figure 5: *ALASKA, , Node 14183 had 1% and 4 occurrences*

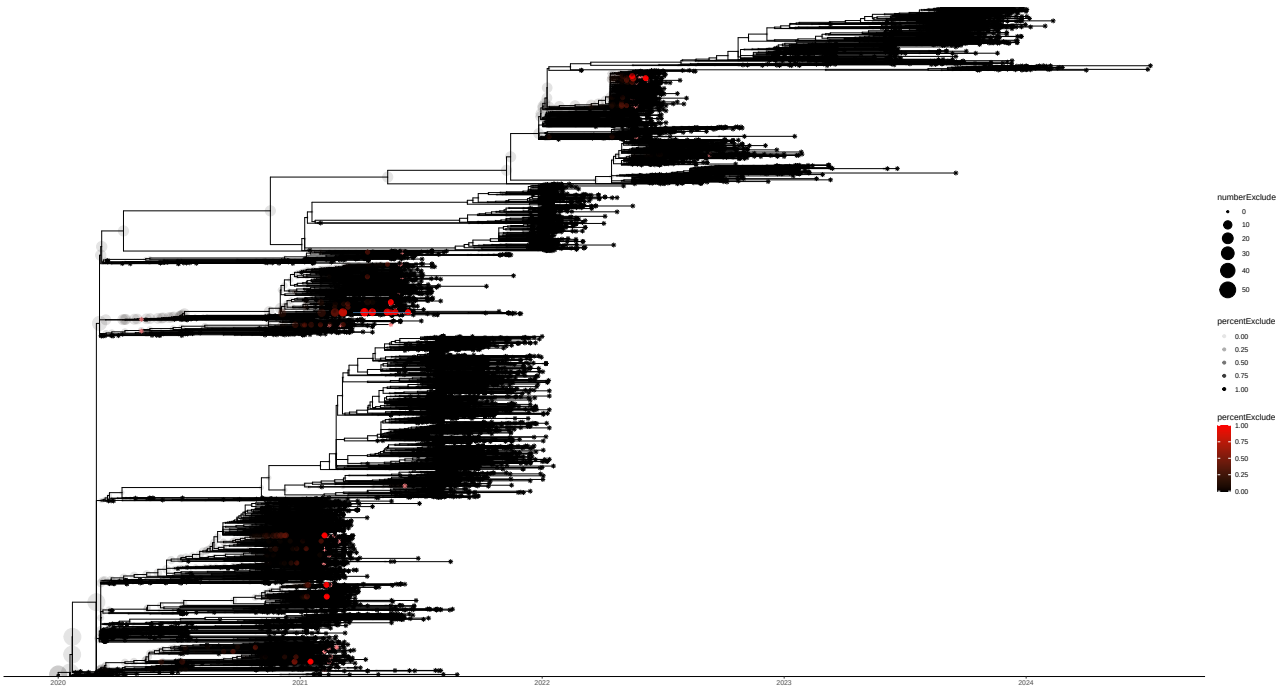

Figure 6: *NEW MEXICO, Node 13978 had 100% and 7 occurrences*

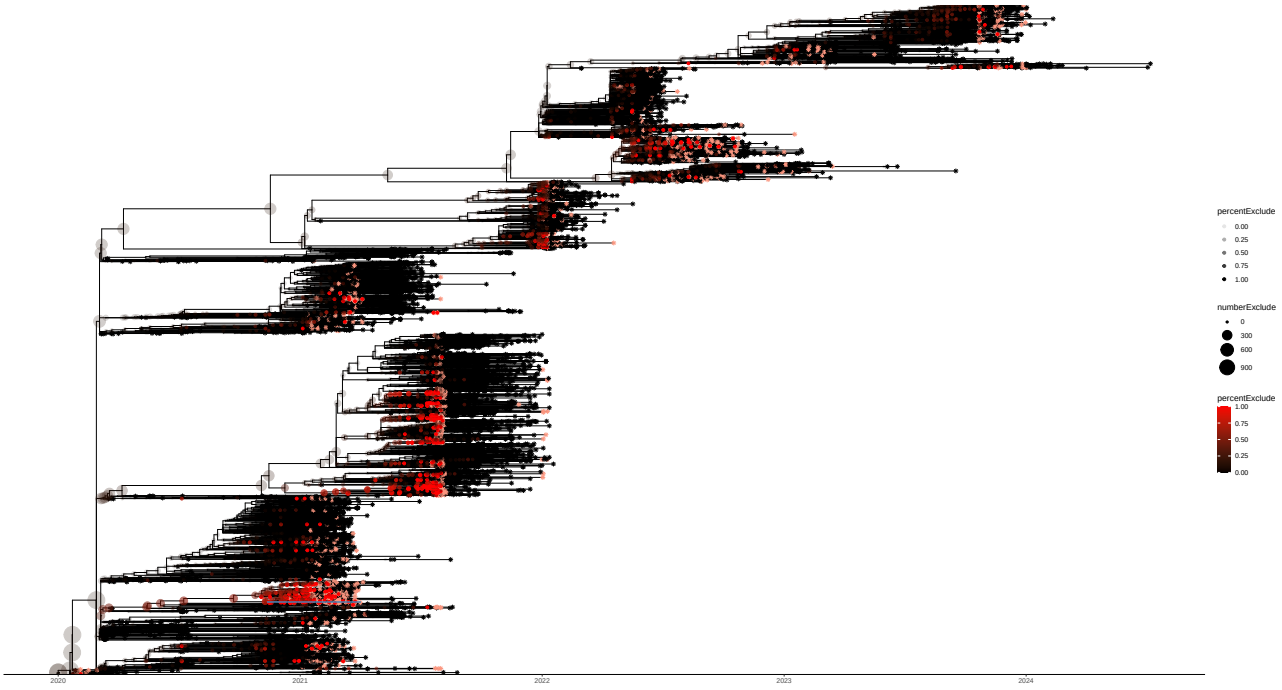

Figure 7: *CALIFORNIA, Node 10302 had 100% and 26 occurrences*

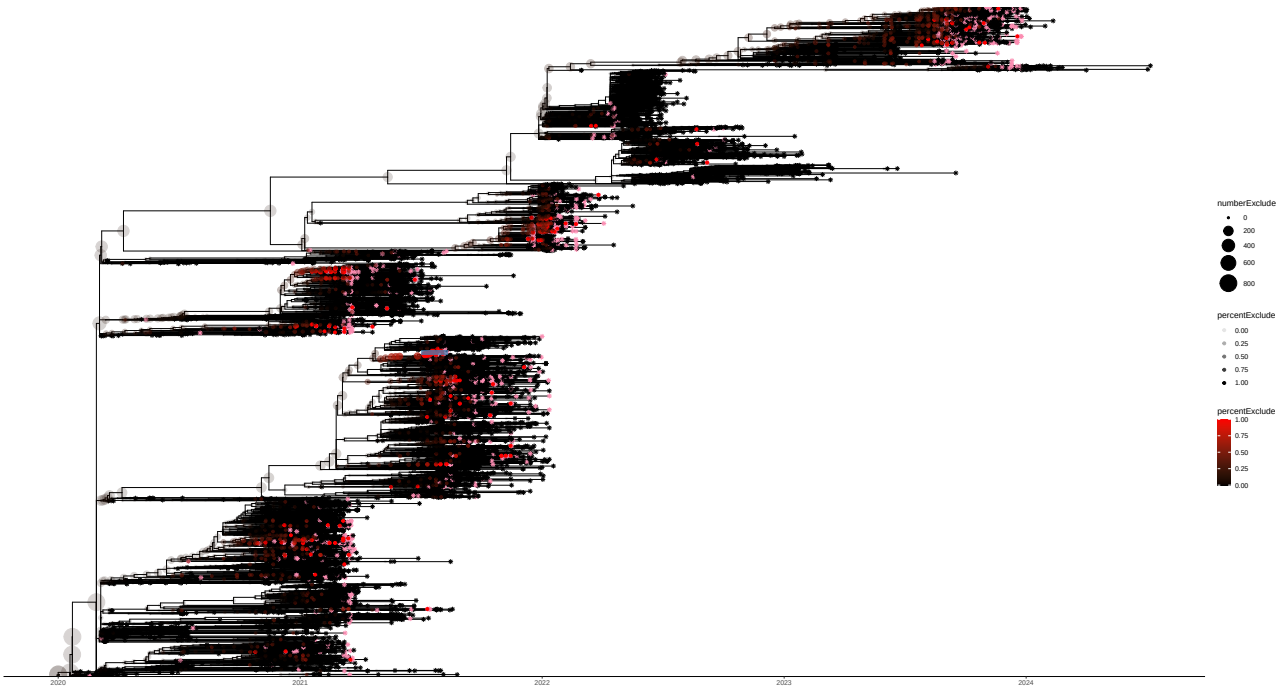

Figure 8: *TEXAS, Node 13462 had 100% and 63 occurrences*

Node 13465 has 100% and 60 occurrences Texas

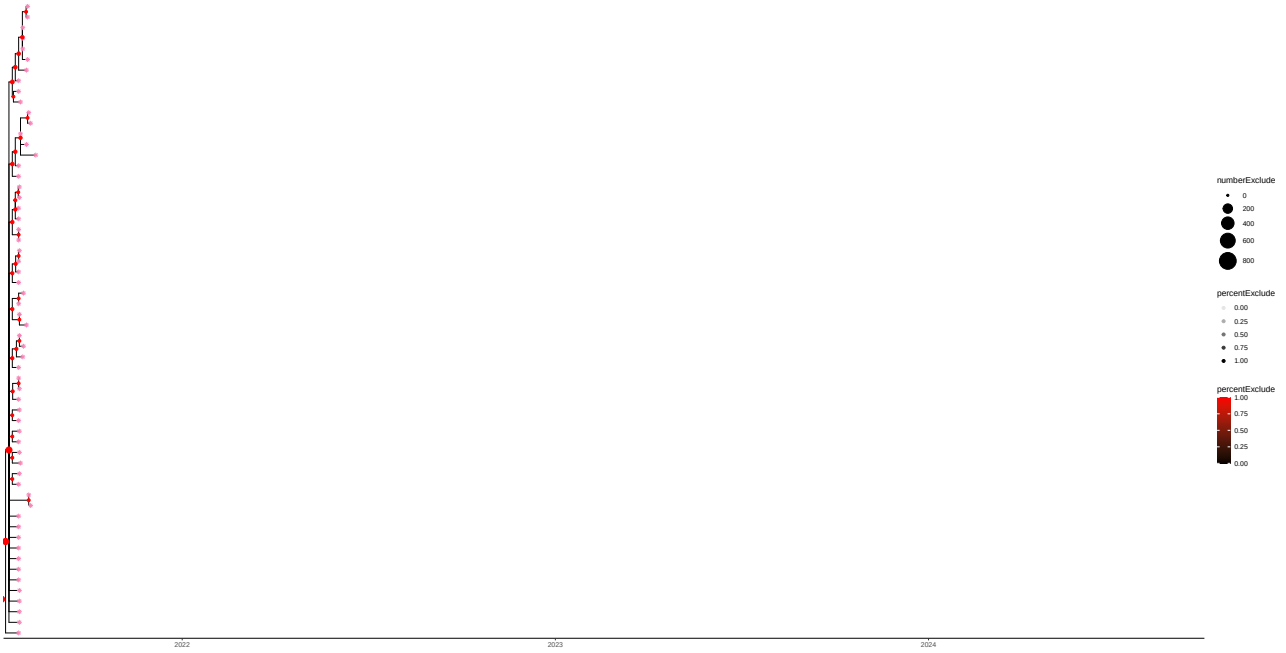

Figure 9: *Texas subtree 13462 had 100% and 63 occurrences*

Figure 9 showcases something extremely interesting. Normally, in a phylogenetic tree, we would expect each internal node to branch off into two distinct nodes. However, Figure 9 demonstrates non-dichotomous behavior with one node branching off into many nodes rather than just two. One reason for why this could occur would be that mutations within the viral genome were not captured fast enough. There most likely, as they always do, exist in-between sequences that are missing from the sequence database due to the host not getting sequenced. Thus, we'd end up with this huge non-dichotomous subtree.

Each state has its own unique set of transmission clusters, but as covered previously, most states fall under categories 1 or 2 of transmission clusters. They simply don't have enough to be significant. Only a couple states stand out as having a significant amount (Figure 10).

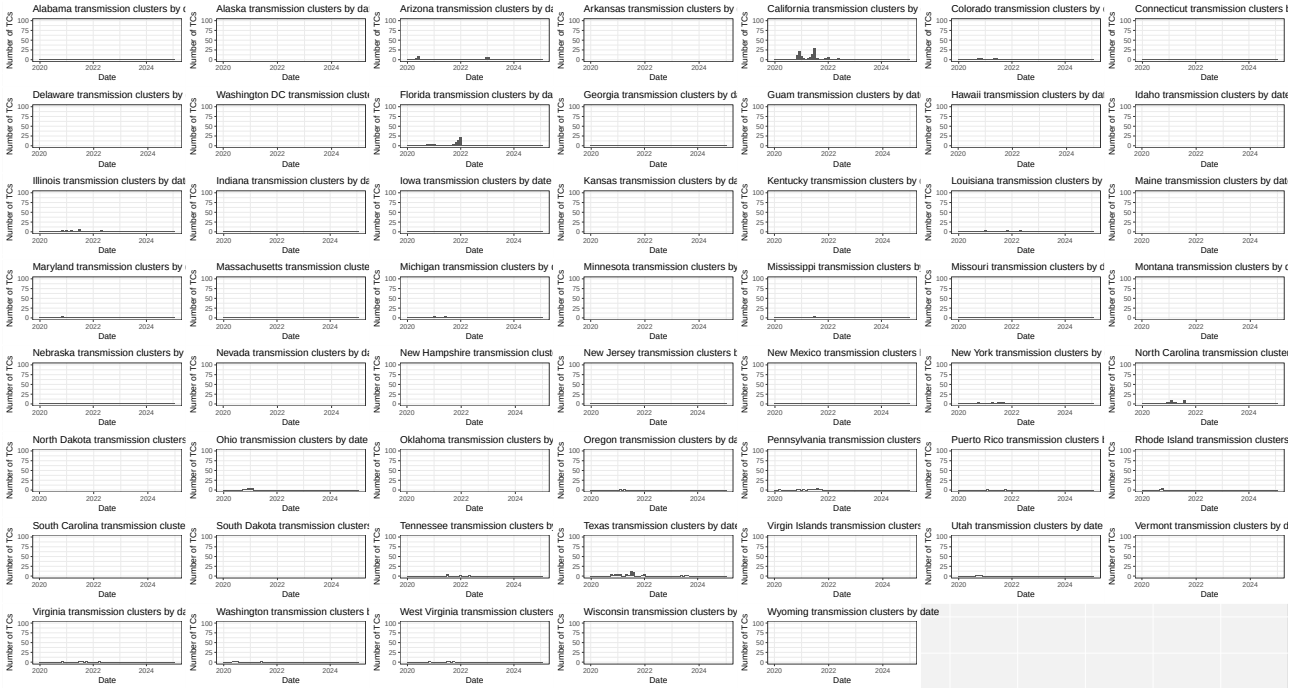

Figure 10: *Cutoff: 50%, 5 occurrences*

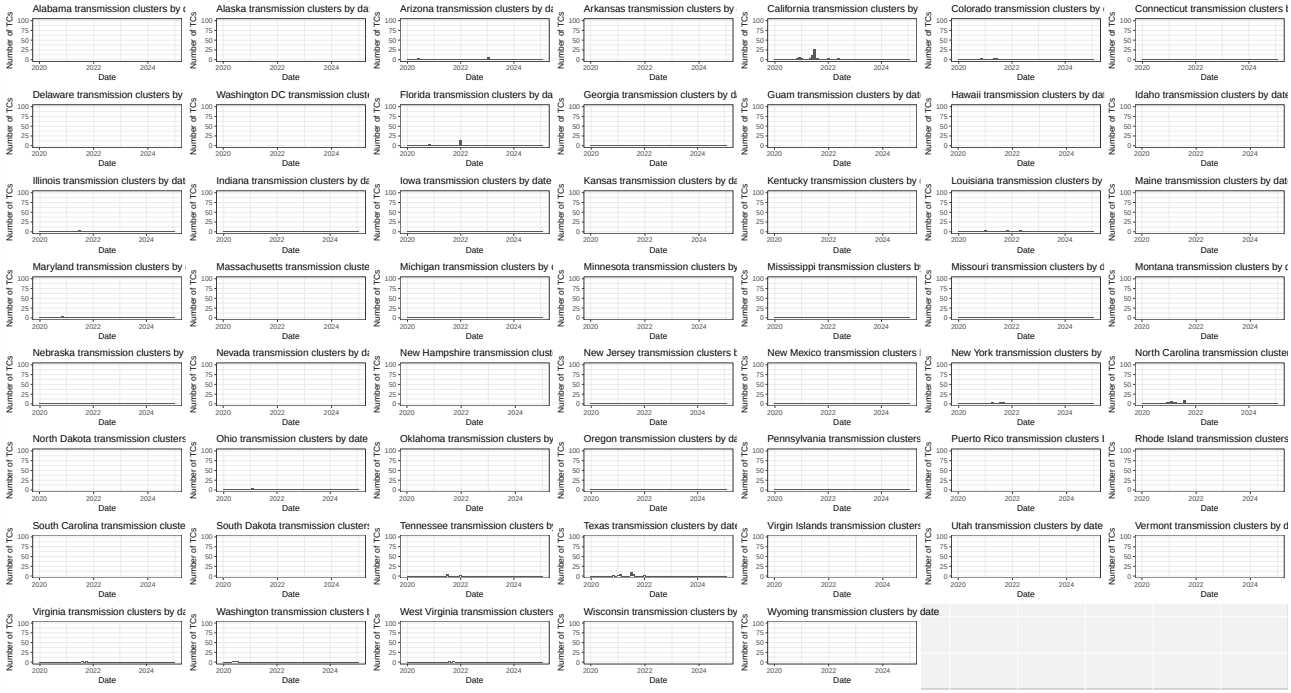

Figure 11: *Cutoff: 80%, 5 occurrences*

If we up the limitation to 80% and 5 occurrences, we see even less transmission clusters (Figure 11). Even though there exist these huge transmission clusters (Figure 9), they are not enough to make up for a lack of other transmission clusters which results in states like Texas not being visible when we raise the cutoff.

5 Further Research Directions

Currently, a widely-accepted definition for transmission clusters has not yet been decided on [6]. One of the reasons for this difficulty is that each dataset will have a different number of locations and thus, determining cutoffs is unique. However, even then, comparing cases by location with different models of transmission clusters could reveal a superior definition of a transmission cluster.

Using ancestral character estimation (ACE) or ancestral state reconstruction instead of the constructed clade percents on each individual internal node would allow a different measurement of transmission clusters.

Jacka et al determined HIV-1 transmission clusters based on co-existing lineages at a particular point in

time. As Hassan et al point out, time could be superior to genetic distance as certain lineages have different rates of mutation [6].

6 Conclusion

A state is more connected than a nation just as a nation is more connected than the entire globe. Bigger cities typically boasted more transmission clusters. However, a subtlety must be made in that bigger cities often had more sequences in general. On the other hand, a state like Texas had less sequences than California within our data, but had much larger transmission clusters.

7 Acknowledgments

I would like to thank the Institute for Computing in Research for providing this opportunity along with my mentor Emma Goldberg for guiding me through this project.

References

[1] "Coronavirus tracker." [Online; accessed August 01, 2024].

[2] O. US EPA, "Indoor air and coronavirus (covid-19)," June 2020.

[3] M. Bousali, A. Dimadi, E.-G. Kostaki, S. Tsiodras, G. K. Nikolopoulos, D. N. Sgouras, G. Magiorkinis, G. Papatheodoridis, V. Pogka, G. Lourida, A. Argyraki, E. Angelakis, G. Sourvinos, A. Beloukas, D. Paraskevis, and T. Karamitros, "Sars-cov-2 molecular transmission clusters and containment measures in ten european regions during the first pandemic wave," *Life*, vol. 11, p. 219, Mar. 2021.

[4] K. Saravanan, M. Panigrahi, H. Kumar, D. Rajawat, S. S. Nayak, B. Bhushan, and T. Dutt, "Role of genomics in combating covid-19 pandemic," *Gene*, vol. 823, p. 146387, May 2022.

[5] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher, "Nextstrain: real-time tracking of pathogen evolution," *Bioinformatics (Oxford, England)*, vol. 34, p. 4121–4123, Dec. 2018.

[6] A. S. Hassan, O. G. Pybus, E. J. Sanders, J. Albert, and J. Esbjörnsson, "Defining hiv-1 transmission clusters based on sequence data," *AIDS (London, England)*, vol. 31, p. 1211–1222, June 2017.