Alayna Daws
Honors Option – STAT 4660
16 May 2022

Phylogenetic Analysis - COVID-19

1. Introduction

COVID-19 is a highly contagious infectious disease caused by SARS-CoV-2 virus. SARS-CoV-2 spread across the world in a very short period. This led to the World Health Organization (WHO) declaring COVID-19 as a global pandemic on March 11, 2020. Most people infected experience mild to moderate respiratory illness (who.int). It is believed to take five to six days, on average, from when someone is infected with COVID-19, to show symptoms. It may take upwards of fourteen days to show COVID-19 symptoms. The most common symptoms include fever, cough, loss of taste/smell, and tiredness. COVID-19 symptoms that require medical attention include difficulty breathing, confusion, chest pain, and loss of mobility. COVID-19 may spread from an infected person's mouth or nose through liquid from breath (who.int).

Coronaviruses (CoVs) are positive-stranded RNA. These viruses have a crown-like appearance under an electron microscope because of spike glycoproteins on the envelope (Cascella et.al 2022). With SARS-CoV-2 being an RNA virus, it is more prone to genetic evolution and mutations over time. Unfortunately, this has resulted in variants that are very different than the ancestral strains.

The transmission of COVID-19 can be reduced and potentially prevented. To slow COVID-19 transmission, individuals may get vaccinated, remain farther than 1 meter from others, wear a face mask in close settings. It is also recommended to wash hands regularly with water and soap. It is important for others to remain at home and self-isolate if infected or feeling ill.

2. Goals

Many SARS-CoV-2 variants have been identified through genetic sequencing. During the duration of the COVID-19 pandemic, there have been five variants considered to be a "variant of concern (VOCs)" by the World Health Organization. These five were identified as VOCs based on their impact on global public health. The five variants are identified as Alpha, Beta, Gamma, Delta, and Omicron. Alpha, Beta, and Gamma are classified as previously circulating. Delta and Omicron are considered to be currently circulating VOCs. Omicron is believed to exemplify a 13-fold increase in viral infectivity and is 2.8 times more infectious than the Delta variant (Cascella et.al 2022).

The goal of this report is to see if the Delta and Omicron variant trends within the data can be identified based on phylogenetic clades. This will be completed by analyzing the

dates the samples were taken and comparing this to the variant more prominent at the time based on the COVID-19 timeline provided by the Centers for Disease Control and Prevention.

3. Data Description

The data was collected from the NIH National Library of Medicine: NCBI Virus in the SARS-CoV-2 Data Hub. One data set was analyzed.

Figures below represent all 80 sequences of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) analyzed under the following conditions: 29,795 nucleotides long, all nucleotides were complete as determined by the NIH NCBI Virus, the geographic region analyzed was the state of Georgia located in the United States of America, the source of isolation of the virus for all sequences was the oronasopharynx. The oronasopharynx refers to the mouth, nose, and pharynx. The data was collected between 03-23-2021 and 02-04-2022.

4. Methods

The R package `phangorn` was used to fit phylogenetic trees from aligned sequence data. The sequence data was imported using the 'read.phydat`command as a FASTA file. A distance matrix was calculated using `dist.ml`. Distance-based methods result in a transformation of the sequence data into pairwise distances. This matrix was then used to fit the tree. The result of the command displays the calculated evolutionary distance between each pair of patients, local controls, and the dentist. Next, an UPGMA tree was computed. UPGMA (unweighted pair group method with arithmetic mean) is a simple hierarchical clustering method from the bottom-up. The UPGMA is based on unweighted averages and is a clustering algorithm based on distances. Distances were calculated between each pair of taxa and these distances are used to construct the tree.

The substitution model named ‘WAG’ was used in the UPGMA tree. The ‘WAG’ model is based on maximum likelihood fits of amino acid substitution data. Next, the maximum likelihood tree was fitted using the `pml` and `optim.pml` functions. The `pml` function takes the UPGMA tree calculated in the previous step as the input and calculates the likelihood. The `optim.pml` function was used to find the maximum likelihood tree. Maximum likelihood is an optimality criterion of character data. The likelihood value for each possible tree was calculated and the best tree was defined as the tree with the highest likelihood. `optNni=T` was specified so that the tree is optimized along the branch lengths and over the topology.

Data sampled from March 2021 to April 2021 were considered to be the Gamma variant. Data sampled from May 2021 to October 2021 were considered to be the Delta variant. Data sampled from November 2021 to February 2022 were considered to be the

Omicron variant. These distinctions are based on the variant that was most prominent at the time as decided by the CDC COVID-19 timeline.

5. Analysis

Figure 1 UPGMA All Sequences

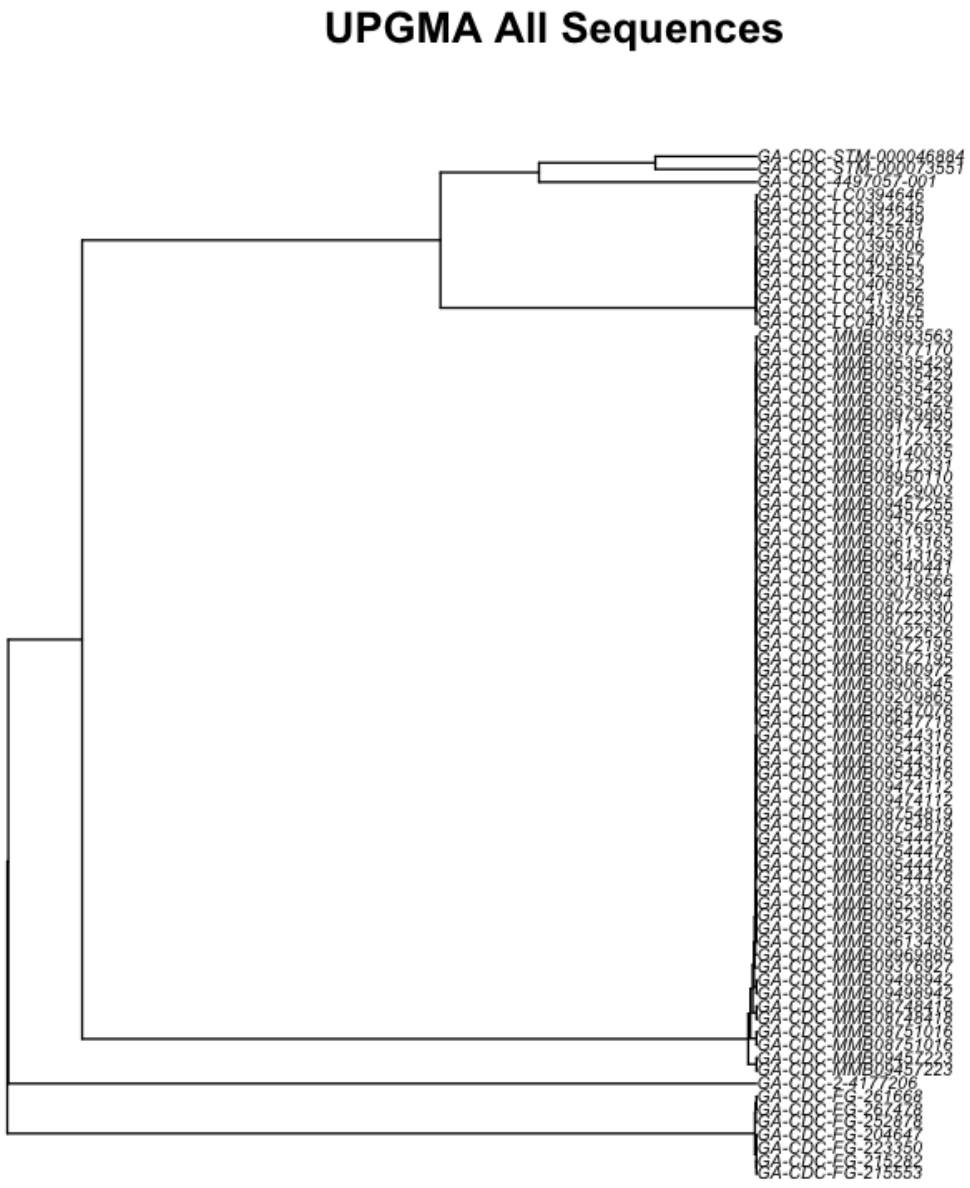

Figure 1 is the phylogenetic tree of all the 80 COVID-19 sequences from the data set of the infected people fitted using the UPGMA. The tree is labeled based on their “Isolate” recorded. It should be noted that multiple sequences are believed to be related as they are within the same clade.

Figure 2 UPGMA All Sequences Fan by Month Group

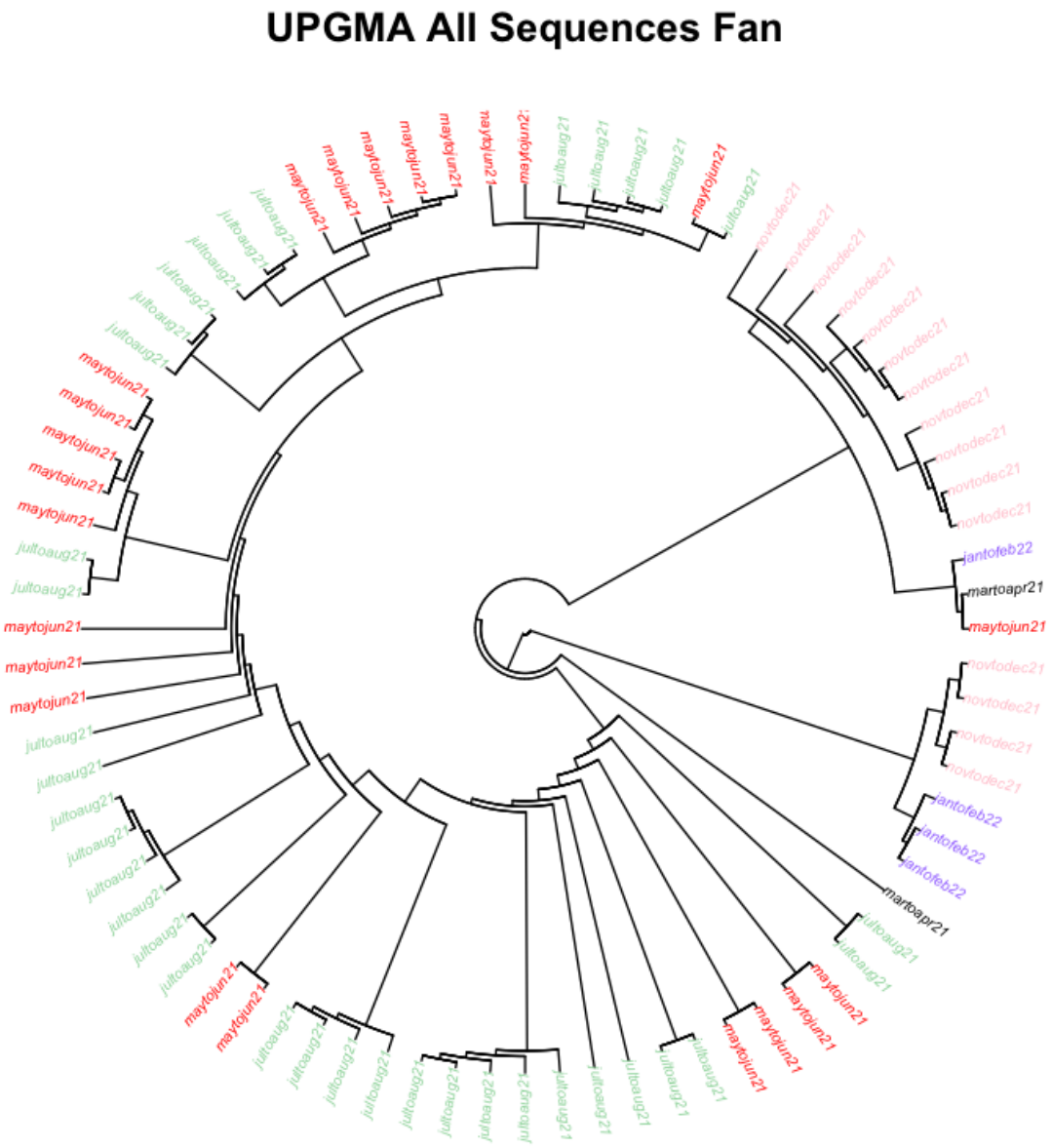

Figure 2 is a fan phylogenetic tree of the 80 sequences from COVID-19 infected people fitted using the UPGMA model. This figure includes all sequences provided in a fan shape pattern. Each person is color coded to the date of their sample collection. This can be used to be able to distinguish between people and determine similar groupings. Typically, the sample collection in the same month groups have similar sequences. The sequences were colored based on their month range and year of the sample collection. March to April 2021 is black, May to June 2021 is red, July to August 2021 is green, September to October 2021 is blue, November to December 2021 is pink, and January to February 2022 is light blue. Many month groups are within similar clades. Month groups close to their corresponding month groups are more likely to be connected to the same internal node.

Figure 3 UPGMA All Sequences by Month Group

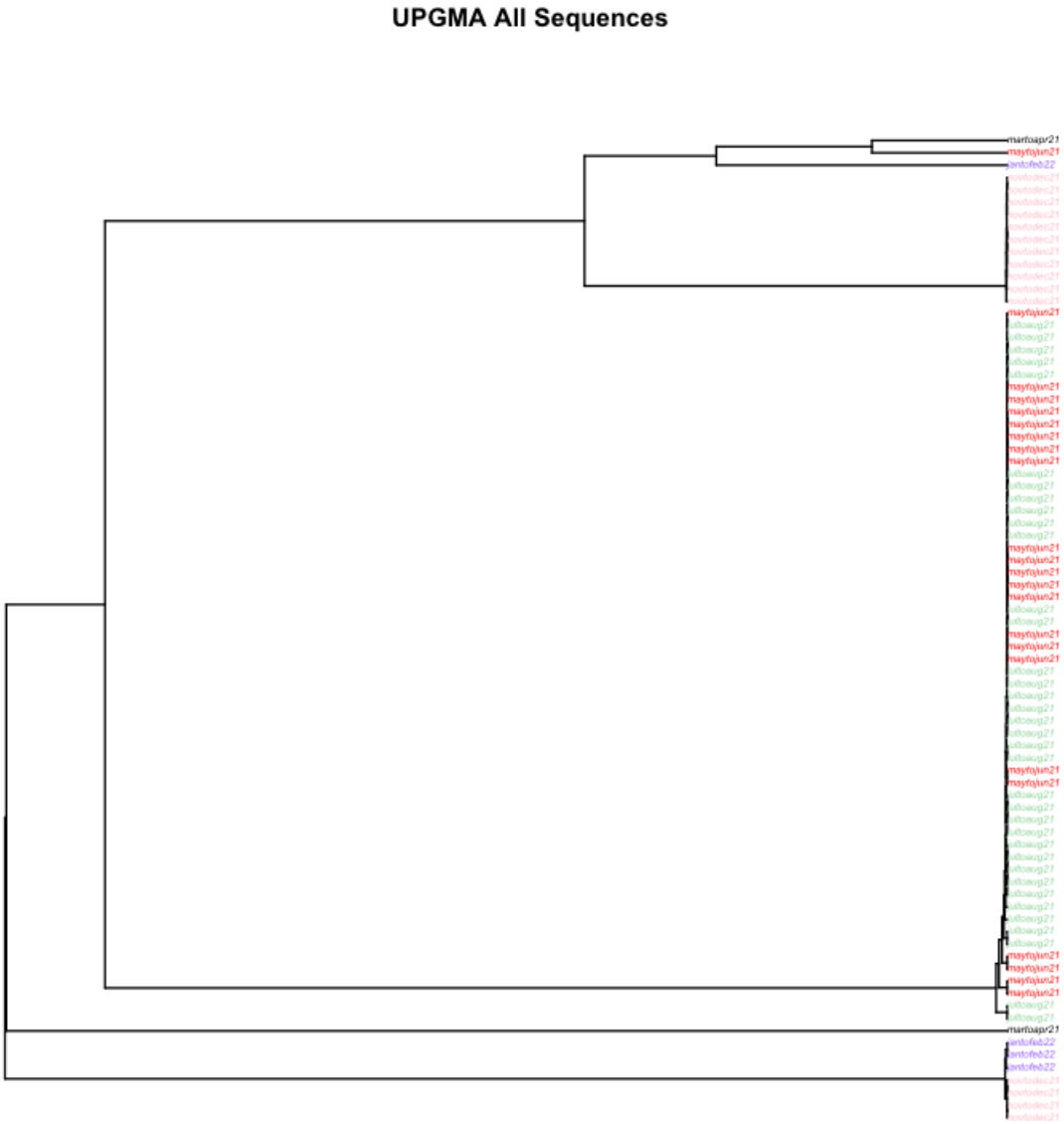

Figure 3

Figure 3 is a phylogenetic tree of the 80 sequences from COVID-19 infected people fitted using the UPGMA model. This figure includes all sequences provided. This portrays the same data using the same model in figure 2. Each person is color coded to the date of their sample collection. This can be used to be able to distinguish between people and determine similar groupings. Typically, the sample collection in the same month groups have similar sequences. The sequences were colored based on their month range and year of the sample collection. March to April 2021 is black, May to June 2021 is red, July to August 2021 is green, September to October 2021 is blue, November to December 2021 is pink, and January to February 2022 is light blue.

Figure 4 – Maximum Likelihood of All Sequences

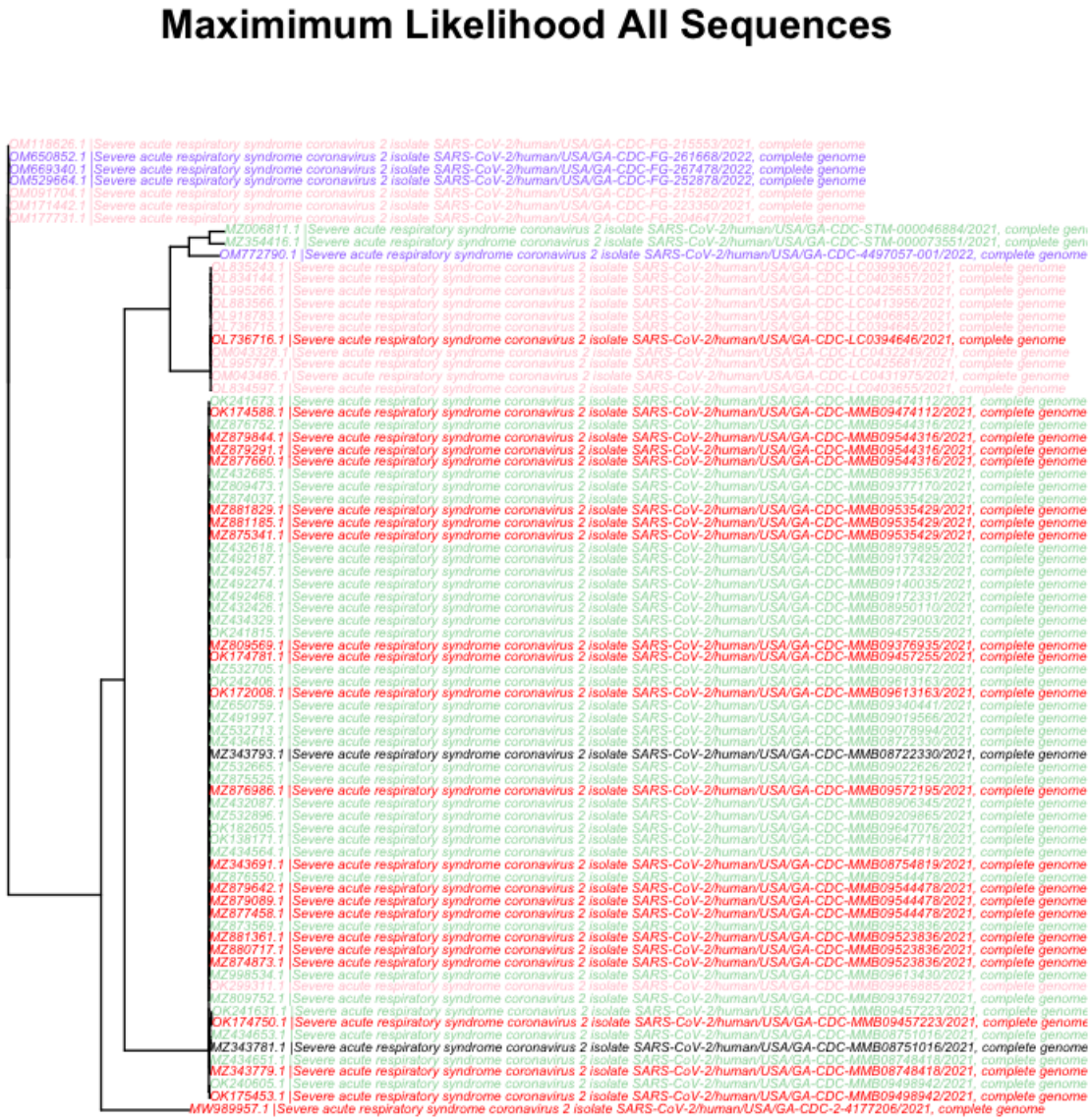

Figure 4 is a phylogenetic tree of the 80 sequences from COVID-19 infected people fitted using fitted using maximum likelihood. Each person is color coded to the date of their sample collection, using the criteria above, with the tip label as their “accession”. The maximum likelihood shows similar color groupings in clusters which may signify a relationship between the date of the sample collection and the COVID-19 sequence similarity. March to April 2021 is black, May to June 2021 is red, July to August 2021 is green, September to October 2021 is blue, November to December 2021 is pink, and January to February 2022 is light blue. May to June 2021 and July to August 2021. Month groupings were sometimes within the same clade. These sequences were likely of the same variant, even if they are in different month groups. See future figures for examples.

Figure 5 UPGMA Sequences in Georgia – Fan Shape

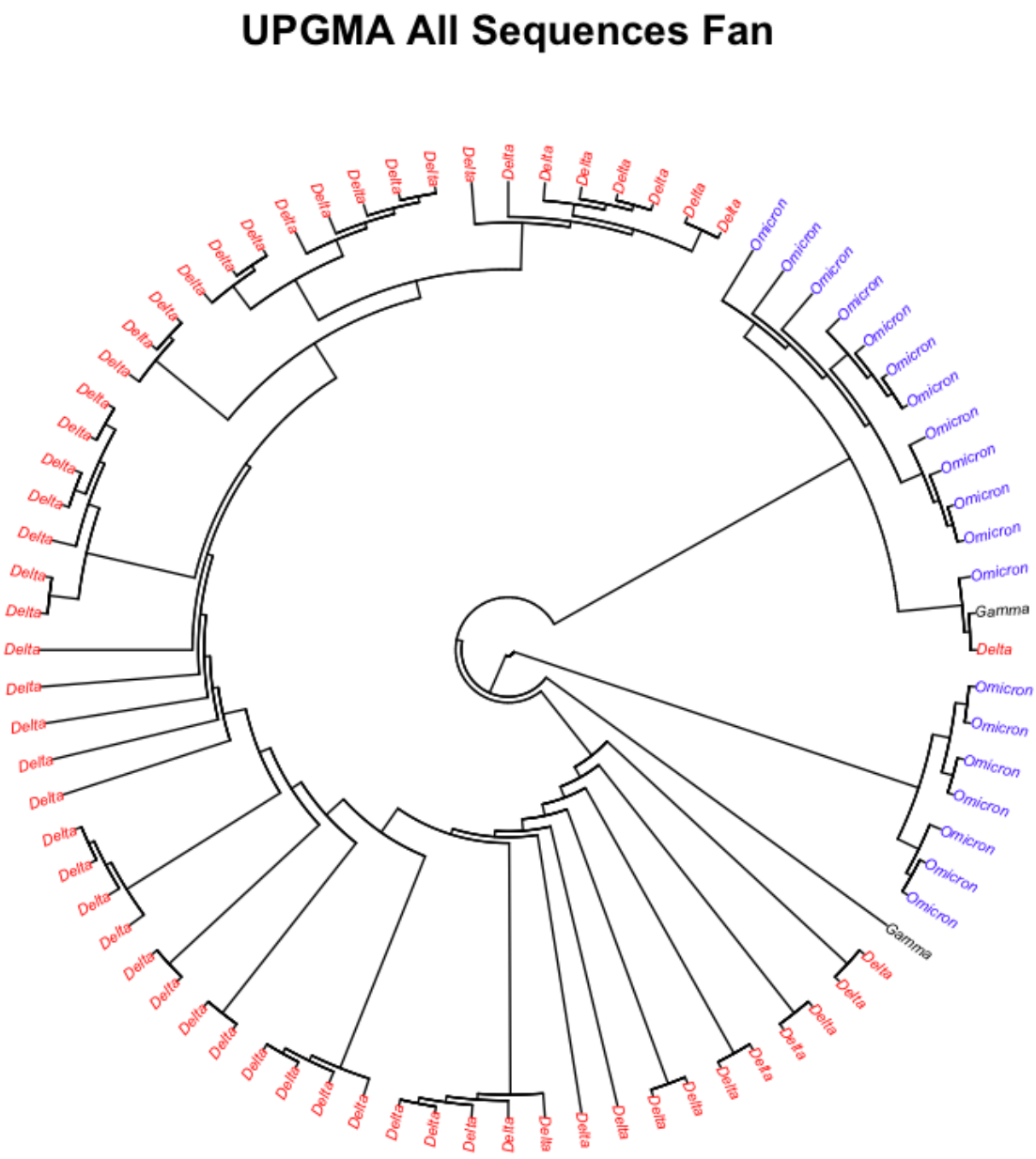

Figure 5 is a fan phylogenetic tree of the 80 sequences from COVID-19 infected people fitted using the UPGMA model. This figure includes all sequences provided in a fan shape pattern. Each person is color coded to the variant that was most prominent at the time as decided by the CDC COVID-19 timeline. If Delta was the prominent variant, then the label is red, Omicron was the prominent variant then the label is blue, and Gamma was the prominent variant then the label is black. This can be used to be able to distinguish between people and determine similar groupings. Typically, the sample collection with the same variant circulating at the same time have similar sequences based on being in the same clade. It should be noted there is an exception with an Omicron, Gamma, and Delta in the same clade.

Figure 6 UPGMA Sequences in Georgia

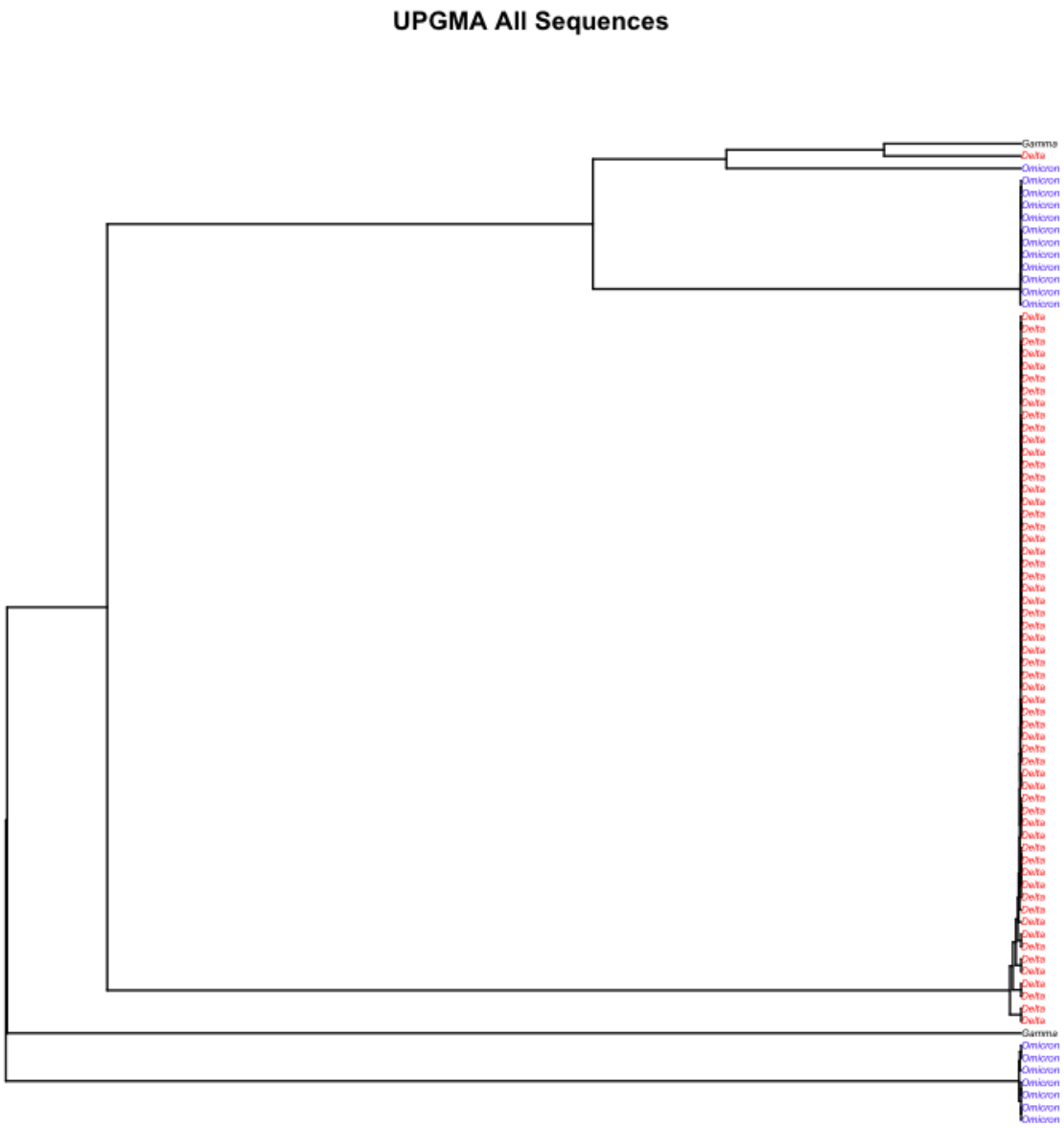

Figure 6 is a phylogenetic tree of the 80 sequences from COVID-19 infected people fitted using the UPGMA model. This figure includes all sequences provided in a fan shape pattern. Each person is color coded to the variant that was most prominent at the time as decided by the CDC COVID-19 timeline. This can be used to be able to distinguish between people and determine similar groupings. Typically, the sample collection with the same variant circulating at the same time have similar sequences. If Delta was the prominent variant, then the label is red, Omicron was the prominent variant then the label is blue, and Gamma was the prominent variant then the label is black.

Figure 7 – Maximum Likelihood of All Sequences

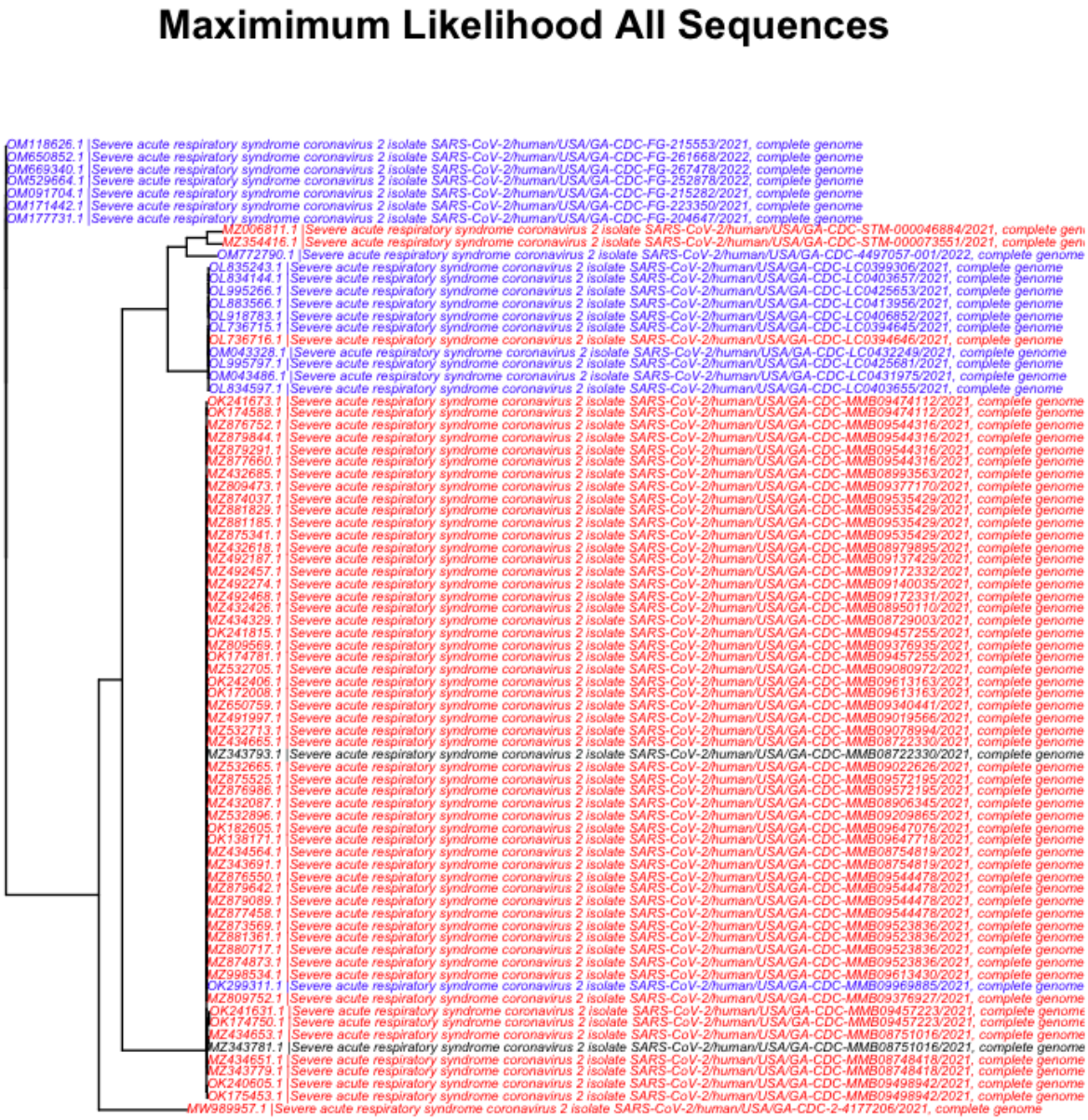

Figure 7 is a phylogenetic tree of the 80 sequences from COVID-19 infected people fitted using maximum likelihood. Each person is color coded to the variant that was most prominent at the time as decided by the CDC COVID-19 timeline. The tip label is their accession. This can be used to be able to distinguish between people and determine similar groupings. Typically, the sample collection with the same variant circulating at the same time have similar sequences. If Delta was the prominent variant, then the label is red, Omicron was the prominent variant then the label is blue, and Gamma was the prominent variant then the label is black.

6. Conclusion

To infer evolutionary relationships between the taxa, the tree must be rooted. The conclusions drawn are based on viral genetic variation. As previously mentioned, COVID-19 is an RNA virus. RNA viruses have a high mutation rate because they are single stranded. As a result, greater variation in the sequences results in more confidence of differentiating the strains.

Based on the analysis, genetic relatedness is observed among the COVID-19 sequences from similar date groupings. These date groupings often relate to the variant more prominent at the time. This relationship was shown through the different phylogenetic trees in the figures above. Similar date groupings often fell within similar clades. Typically, the sample collection with the same variant circulating at the same time had similar sequences.

The Delta and Omicron variant trends within the data were identified based on phylogenetic clades and phylogenetic relatedness. These methods can be applied to future variants to visualize variant relatedness.

7. Assumptions

UPGMA is not a good model when the substitution rate varies between branches. The substitution rate should be considered when drawing conclusions from the UPGMA model. If the substitution rate is high, a different distance method should be used. The maximum likelihood approach results in a very large number of possible trees when the sample size increases. As a result, maximization may become too difficult. Computational time and costs required may be too great in the maximum likelihood approach.

8. References

Sources:
https://www.who.int/health-topics/coronavirus#tab=tab_
https://www.ncbi.nlm.nih.gov/books/NBK554776/
https://pubmed.ncbi.nlm.nih.gov/27392606/
https://www.yalemedicine.org/news/covid-19-variants-of-concern-omicron
https://www.cdc.gov/museum/timeline/covid19.html