Visualizing Covid Spike Protein Mutation Over Time

Vincent Lombardi: 11731142
Ajeigbe Oluwafemi: 11771336
Pierce Cappa: 11570387

Abstract

The COVID-19 crisis has been one of the most talked-about as it has negatively impacted human living over the past months. The etiology of this disease is the severe acute respiratory syndrome coronavirus 2 (SARS COV 2). The spike protein of the SARS COV2 is what the virus uses in receptor recognition and the cell membrane fusion process. The primary goal of the COVID vaccine was to create neutralizing antibodies that would fight the most critical part of the SARS COV 2 virus; the spike protein. Recent studies on the SARS COV 2 virus have observed that the virus has been found to mutate in human-to-human transmissions over time, and these changes can potentially alter the efficacy of proposed treatments. This project aims to analyze the sequence and structure data for the COVID-19 spike protein, explore variations and mutations in the spike protein over a given timeframe, and explore the frequency of these spike protein sequences by country. We collected spike protein sequence data from the GISAID [9], which was mapped against the reference data from the NCBI [8] databases for this project. We used R and Python to clean, explore, and visualize the temporal and spatial evolution of the spike protein sequences. While the increasing mutations are noted in other research, our findings also show that the number of mutations accelerates over time.

Intro

Covid-19 has had a dramatic impact on our global society and its effects will be felt for decades to come. After emerging in Wuhan, China, the virus spread across the world in a matter of months. It is arguably one of the worst pandemics in human history. It also led to widespread stay at home orders and severely hurt the economy.

The virus's success can be attributed to its long incubation period, infectiousness, and a high percentage of asymptomatic carriers. Asymptomatic patients are also a rather significant concern as the virus is nowhere near as deadly or even active in young people than it is in the elderly. This means that it can spread undetected through a rather large group of people before emerging a few days later.

The virus's rapid rate of mutation has caused particular concern as several new variants have popped up since the global pandemic started. This allows the virus to rapidly adapt and overcome certain treatments or simply become better at spreading itself. Some newer variants have been shown to affect younger people more than the original variant. Examples include the Delta variant which was even more infectious than the original variant. Other variants have been shown to be more likely to break through the protection offered by our current vaccines.

The spike protein is what the virus uses to penetrate the cell membrane. By using the spike protein to bind to receptors on the cell's surface it tricks the cell into allowing the virus to fuse to its membrane. The virus then releases RNA causing the cell to produce more viruses. The spike protein is also one of the main targets for vaccines particularly the mRNA ones. Specifically, the Pfizer and Moderna vaccines have been known to target the spike protein.

According to Huang [1], The spike protein can be divided into two different subunits, S1 and S2. The S1 subunit contains the N terminal domain (NTD) and receptor-binding domain (RBD). The S2 subunit contains the fusion peptide (FP), heptapeptide repeat sequence 1 HR1, HR2, TM domain, and cytoplasm domain (CT). The RBD domain is responsible for interfacing with a target cell's membrane. HR1 and HR2 are responsible for regulating the RBD. These mutations can either increase the spike protein's effectiveness or make it harder for current vaccines to target.

Problem Definition

Our project specifically aims to explore covid-19's spike protein mutation over time. This will be done by examining specific portions of the spike protein as well as examining the whole spike protein sequence. Through our exploration, we aim to examine what portions of the spike protein seem to be mutating and when as well as what this could mean? Is the spike protein mutation accelerating over time or is it mutating consistently and at the same rate. In what countries are the most mutations occurring in and why could this be.

These are important questions to answer as for one, each portion of the spike protein is important when examining the effectiveness of the disease but also the effects of said disease. Most portions of the spike protein have been labeled and documented and have their own unique functions and purposes. Thus understanding what portions of the disease have been mutating can allow us to better predict and react to further mutations through vaccines or other governmental policies

The other purpose behind this project is to examine the rate at which the virus is mutating over time. Understanding if the virus is mutating more or less overtime is important information that can be used to inform governmental policies. The portions of the virus mutating can also inform what areas need to be invested in as if only one portion of the spike protein is mutating. The effect of any virus mutating is that new vaccines need to be developed to counter the said virus. Specifically, there have been many questions about the effectiveness of current vaccines on the omicron and delta variants. These new variants have spurred the development and recommendations for covid vaccine boosters. If the rate of new variants appearing is increasing, than it would follow that the development

Models/Algorithms/Measures

To complete this project, one of the first things we had to identify was what algorithms we needed to successfully categorize the spike protein mutations. One of the first types of algorithms needed was a distance algorithm to calculate the distances between different strains. This distance algorithm was needed to implement a number of models that used hierarchical clustering models. The models in question were dendrograms and Treemaps. Because the spike protein sequence was represented as a string in our data set, we specifically needed a string distance algorithm. During our research, we found many algorithms that would have accomplished this task but we decided on using the Damerau-Levenshtein algorithm for both R and Python distance calculations. We used this algorithm in a couple of instances, the first was in calculating the average distances between the different parts of the spike protein over time. This was done specifically to the HR1, HR2, NTD, FP, TM, and RBD sections of the protein sequence. This calculation was also done to the sequence at large to map out the most likely parent sequences where the parent sequence was the sequence with the smallest distance that was documented before this sequence.

Related Work

Temporal changes in spike protein mutations have also been examined in related research. For instance, Shiyu and Samuel (2021) [5] examined the COVID-19 spike protein sequences using statistical approaches to analyze the spike protein sequence mutations and their 3-D structure data. The first approach in this study explored the evolution of the spike protein sequence using the Bayesian hierarchical model to study these mutations' spatial and temporal patterns. The second approach was the sampling algorithms approach to check the possible changes in the 3-D structure of the spike protein. Shiyu and Samuel used the spike protein mutation type called the D614G and four others to examine how each key position of the amino acid type of the spike protein

has changed over time. The transition of the positions 614 from D to G of the D614G sequence mutations increased over time across regions in the world compared to the reference data. Shiyu and Samuel reported weak evidence to support the theory that the most common sequence mutations observed per cluster were due to changes in the 3-D structure of the spike protein near the mutation location.

Jun Lan et al. 2020 did a study on the SARS COV 2 spike protein binding process by analyzing the crystal structure of the spike protein's receptor-binding domain (RBD). Reports from this study presented that the best way to fight the virus was to understand the initial step of infection at the atomic level. This was achieved by comparing the interactions at the SARS COV2 RDB-ACE2 with an earlier version of the virus; SARS COV RBD-ACE2. One notable finding the study reported was the similarity in terms of the networks of hydrophilic interactions of the RBD-ACE2 interfaces. However, further research by Jun Lan et al. reported that a difference discovered between the two viruses could be the unique furin- cleavage site between the two subunits(S1, S2) of the SARS-CoV-2 spike protein, increased the rate at which it infects, suggesting that the SARS COV 2 virus has evolved. Lizhou Zhang et al. 2020 has also done a similar study on the SARS COV2 spike protein. The study examined the D614G mutation of the spike protein and its rate of infection. The analyses of the spike protein sequence variation showed increased infectivity compared to other COVID variants.

Implementation

We used the GISAID global covid spike protein data set. It contains. Roughly has 4808593 entries. Not every entry was used due to either having incomplete sequences or errors in the header that made it hard to determine when or where the sequence was from. The data set contains sequences from a variety of different locations around the world including the United States, UK, France, China, Canada, Brazil, Bangladesh, and many others. It did not contain information on where the sequences were specifically from in said countries so we do not have the specific region. We also changed the dates in the data set to be the first day of the month for ease of use. We removed any sequence that did not have a complete date as we could not be sure when it was collected.

Some of our limitations are that the dataset seemed to be heavily biased towards the United States and the United Kingdom. On the other hand, we had fewer cases from China which means that our data set is unbalanced. Another limitation is that we do not have the exact figures for how effective each mutation is. We can make an educated guess based on the number of times we see it in the dataset but there could be other reasons for a sample appearing a lot. For example, older variants will logically have

more entries in the dataset than newer ones despite some of the newer ones being reportedly more virulent.

Our implementation mainly used R and Python. We cleaned the data with python's pandas library. We removed incomplete sequences as well as sequences that did not occur more than a handful of times. It should be noted that we were not able to get that much data from China as many of their entries had a different header structure which prevented us from easily identifying them. Since most of our analysis was done over time this was not a major issue.

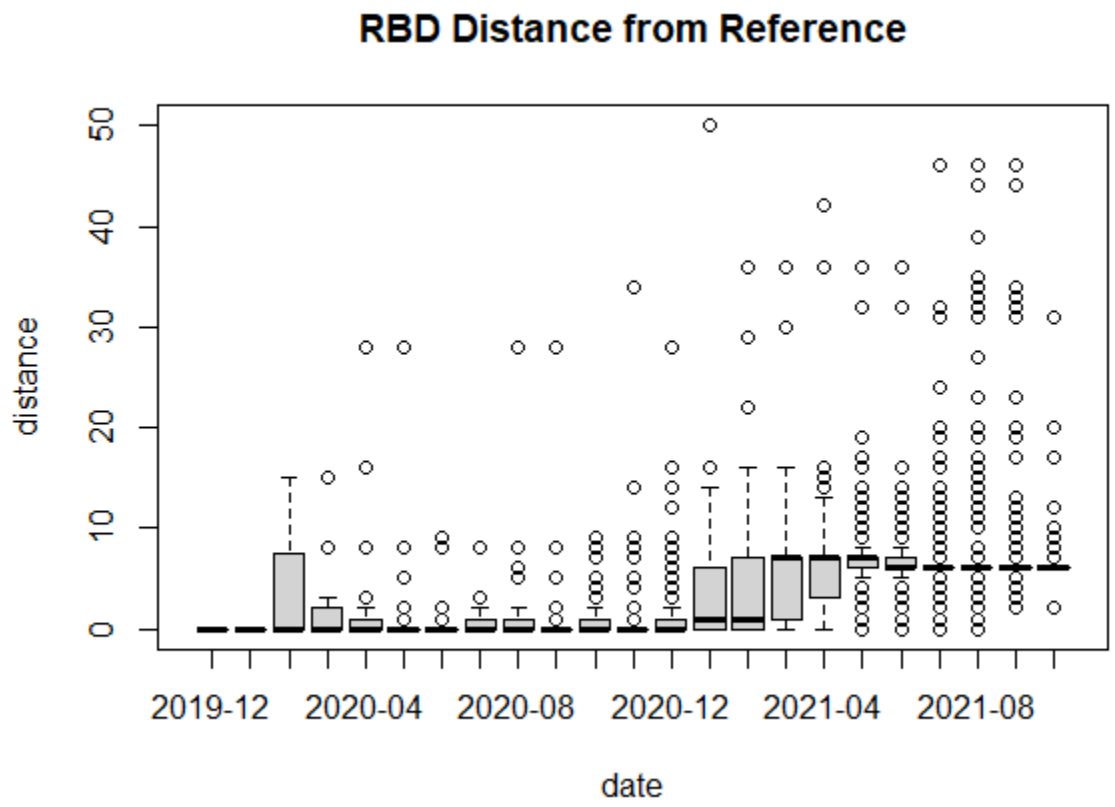
We grouped each of the entries by time and likeness. Essentially we counted the number of times each sequence occurred in a month. This gave us an idea of what strains were more prevalent. We removed repeats and those strains that were below 10 repeats per in any given month. While we thought about using the total number of occurrences we found that it skewed the data too far towards older sequences which were less than informative. We then analyzed the proteins by graphing the subunits in R. We used a boxplot for the graph with an X-axis of distance and a Y-axis of time in months. The distance was measured as Levenshtein distance from the reference sequence.

Using the above methods as a starting point, we also created and organized a second list in python that was used for categorization. This involved grouping each strain by the most closely related based on string distance strain that was documented prior to each given strain. To accomplish this we first cleaned the dates of each strain and then organized the remaining data based on the time of when the covid spike protein variant first occurred from earliest variants to latest variants. We then used the resulting data to map parent and children relationships by comparing a given variant's string distance to the distance between all previous strains in the list and marking its parent as the strain with the smallest string distance.

We were then able to calculate a given strain's success based on how many times the individual times' strain was recorded, as well as how many times a strain's children were recorded. The algorithm used to calculate strain's successes would first create a new column in the data frame and set each value to the given strain's number of times recorded. It would then go through the data frame starting from the bottom and ending at the top, adding a given row's success to its parents' strain success count.

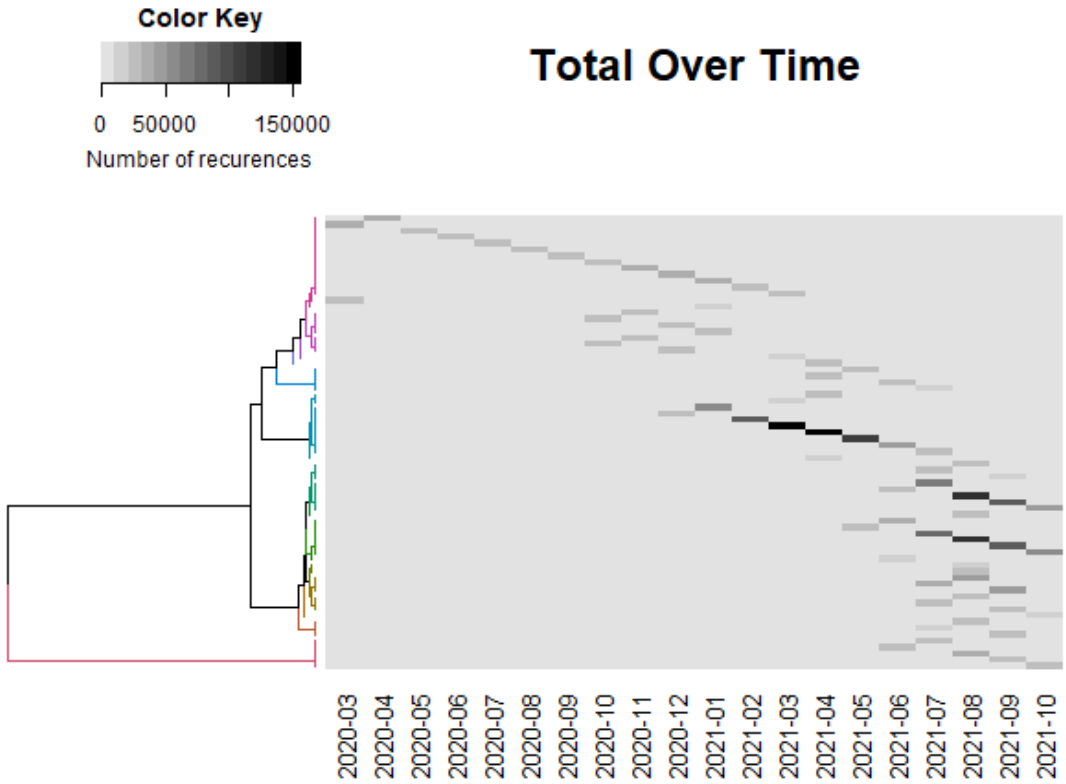
Results

We saw a lot of RBD mutations in early 2020 and early 2021. This makes sense as it plays a critical part in attaching the virus to a target cell's membrane. NTD underwent major mutations in early 2020 and then stopped for a bit before ramping up in 2021. Interestingly HR1 and HR2 both began experiencing major mutations at the beginning of 2021. We do know that the Delta variant emerged in late 2020 so this could have been the start of it. FP, CT, and TM, also underwent some major mutations at the beginning of 2021. The fascinating part is that FP, CT, and TM all underwent major mutations around the same time. We can only assume that these are part of why the delta variant would be successful although admittedly this is only a correlation.

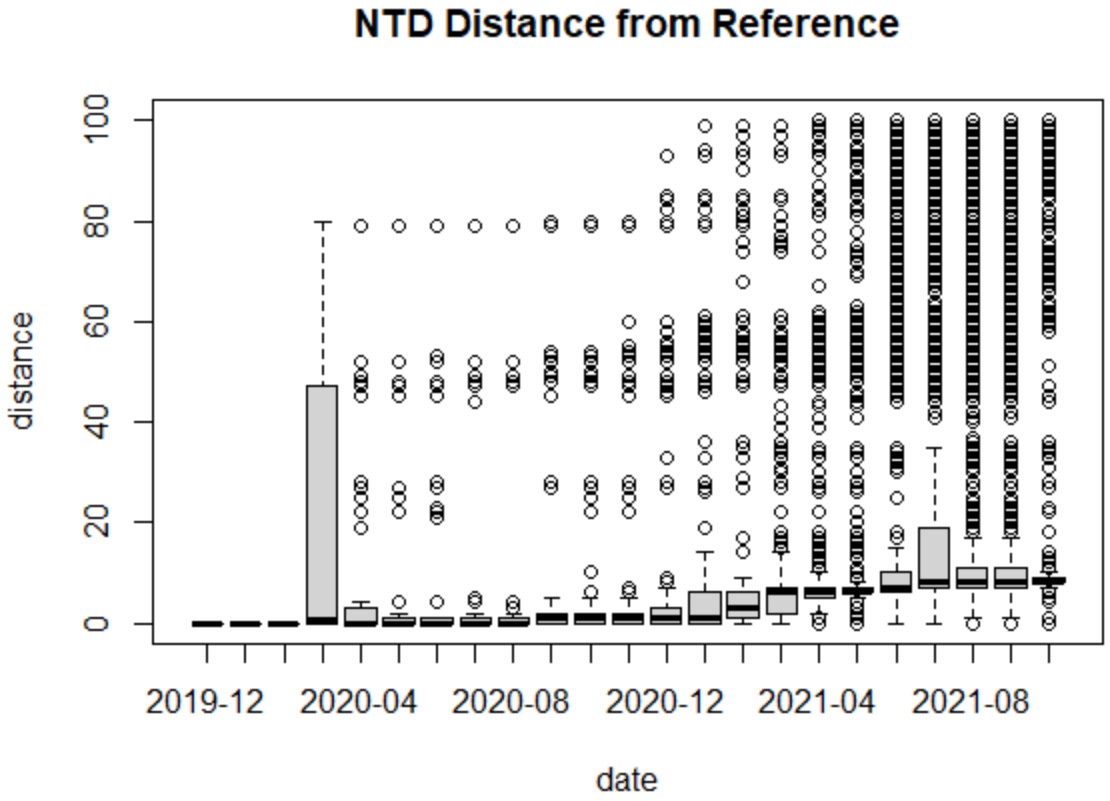


The Total Over Time or mutation or total mutation graph shows that there was a significant uptick of mutations or recorded sequences beginning in 2021. The total over time graph shows that there was a dramatic increase in mutation variance near the beginning of 2021 as shown by the fact that the heat map becomes more spread out. The Total over Time graph is a Dendrogram where the top-performing strains for each month are collected together. The sequences are grouped together with their close neighbors so you can get a rough idea of how the virus mutated over time. The limitation of this graph is that it can only display so much. The reason we had to go with

monthly recurrence vs total recurrence for filtering was that otherwise the graph never showed newer mutations. The other main limitation is that it does not show the subunits so at most you can only look to see the general dispersal of active variants.

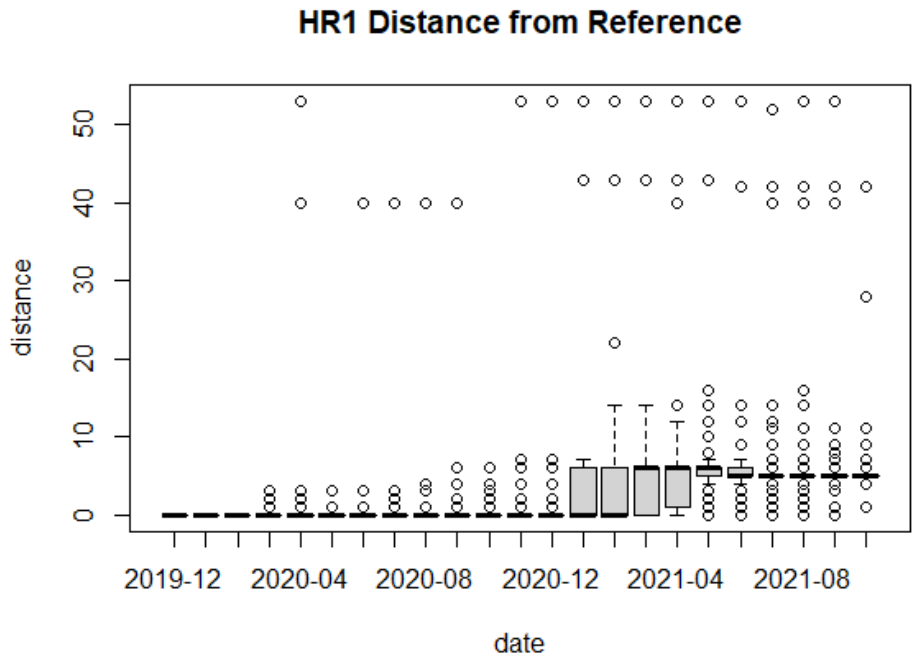


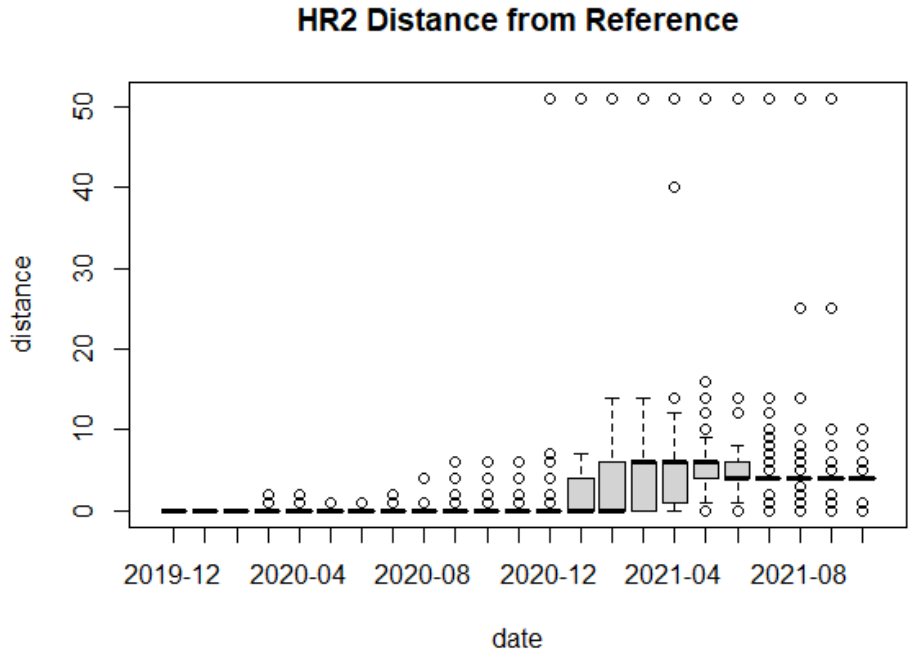
Overall there were two periods of high mutation for the spike protein. One in spring 2020 and one larger one in summer 2021. This coincides with spikes in covid cases around the world in late 2020 and early 2021 Ritchie [2]. Summer 2021 was not as bad as 2020 but this could have been due in part to the increase in vaccinations. What is telling is that despite the rise in vaccination rates during the early summer of 2021 Covid cases still shot up in late 2021. This was especially true in the UK which had high vaccination rates but high infection rates. I would say that the mutations in the spike protein likely are the cause. It is difficult to determine what specific mutation would have caused this uptick in cases for summer 2021 as nearly the entire spike protein underwent changes during this time period.



One other observation is that there is not a lot of variance in the data set after early 2021. As in most of the recorded sequences are not in the data set. This could either be that fewer reference sequences were reported in later 2021 or maybe it's a sign that the original virus is being out computed by newer strains. The odd part is after the first RBD mutation we do not see a large number of mutations until 2021. This can be seen in both the HR1 and HR2 graphs below. Both graphs also had noticeable similarities in the rough placement of their average distance.

We expected there to be more variance between the different protein components but with the exception of NTD and RBD, they were all similar. It should be noted that the graphs are not scaled so the specific numbers on the Y-axis differ between subunits as they are all of different lengths. That being said you can still see the same pattern in nearly all of the graphs. This is also why we do not show the graphs for TM, CT, and FP as they look very similar to HR1 and HR2.





Some of our limitations are that the dataset seemed to be heavily biased towards the United States and the United Kingdom. On the other hand, we had fewer cases from

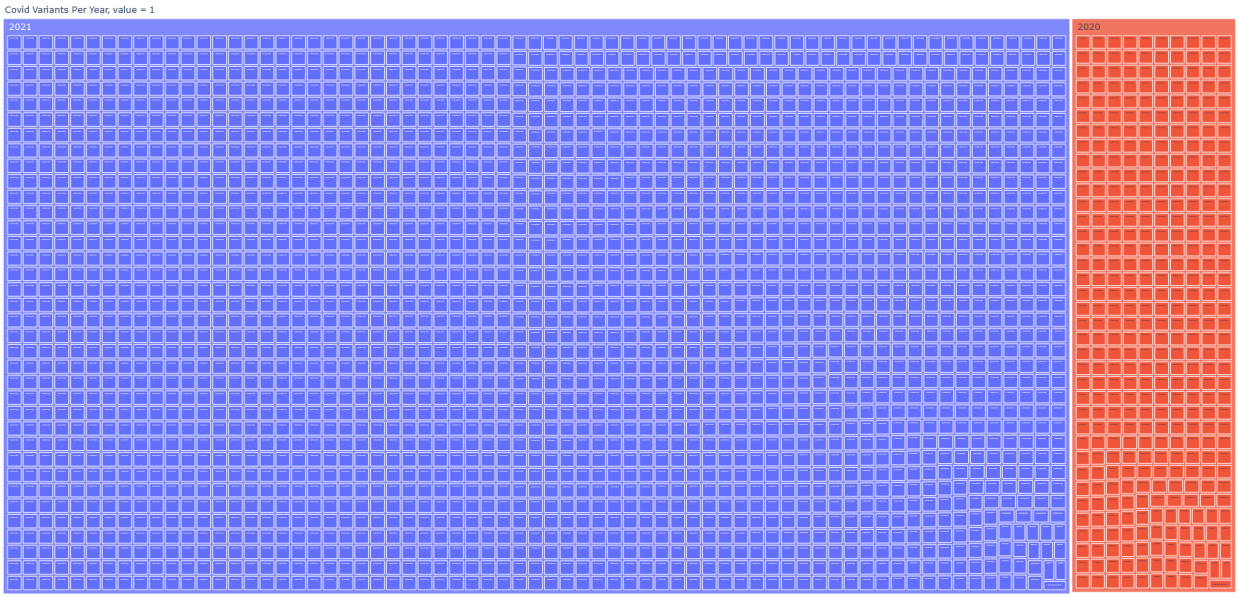
China which means that our data set is unbalanced. Another limitation we have is that we do not have the exact figures for how effective each mutation is. We can make an educated guess based on the number of times we see it in the dataset but there could be other reasons for a sample appearing a lot. For example older variants will logically have more entries in the dataset then newer ones despite some of the newer ones being reportedly more virulent.

One of the visualization techniques we also did was group the sequences by country and frequency. The world below was designed using the leaflet package in R. The goal of this visualization was to create an interactive map that plots each sequence count of the SARS COV2 per country. On the map created, the color of dots in countries represents the number of new strains found, where red represents a low number of new strains and black represents a high number of new strains. We found that the USA has the most significant number of new sequences, most likely caused by the amount of testing and frequency of new cases. Germany also had a high number of new sequences found.

Figure: Sequences grouped by Country

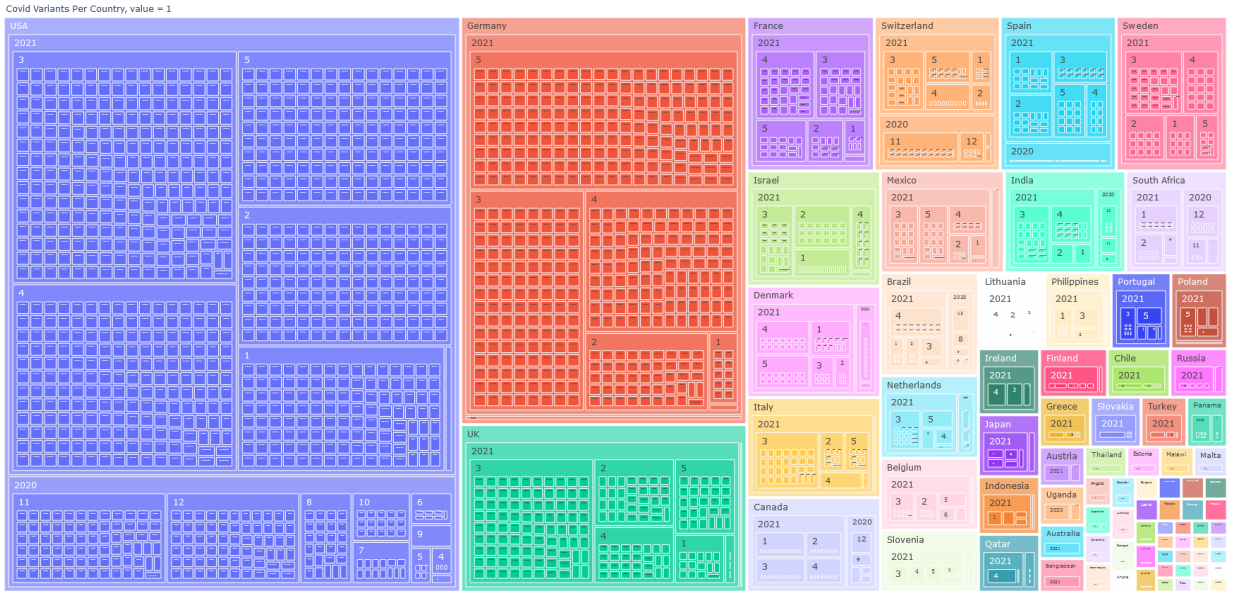
We furthermore explored the mutation of covid-19 mutations using tree maps using the second dataset where each mutation was paired with its parent and strain success. In these treemaps each unique strain was plotted alongside one another and

placed into different groups depending upon path. The path of each node could be as simple as the year in which the strain was first recorded, as well as country, parent strain, and then strain itself. The size of each box was set to either strain success, calculated as described above, or set at 1 per box which would emphasize the number of new mutations rather than how prolific these mutations had become. Another thing to note is that the data used covers 2019, 2020 and the first half of 2021. Through these diagrams a number of key observations can be made. The first is that it seems clear that the number of covid mutations is accelerating, and accelerating quickly at that. This can be observed in the below tree map.

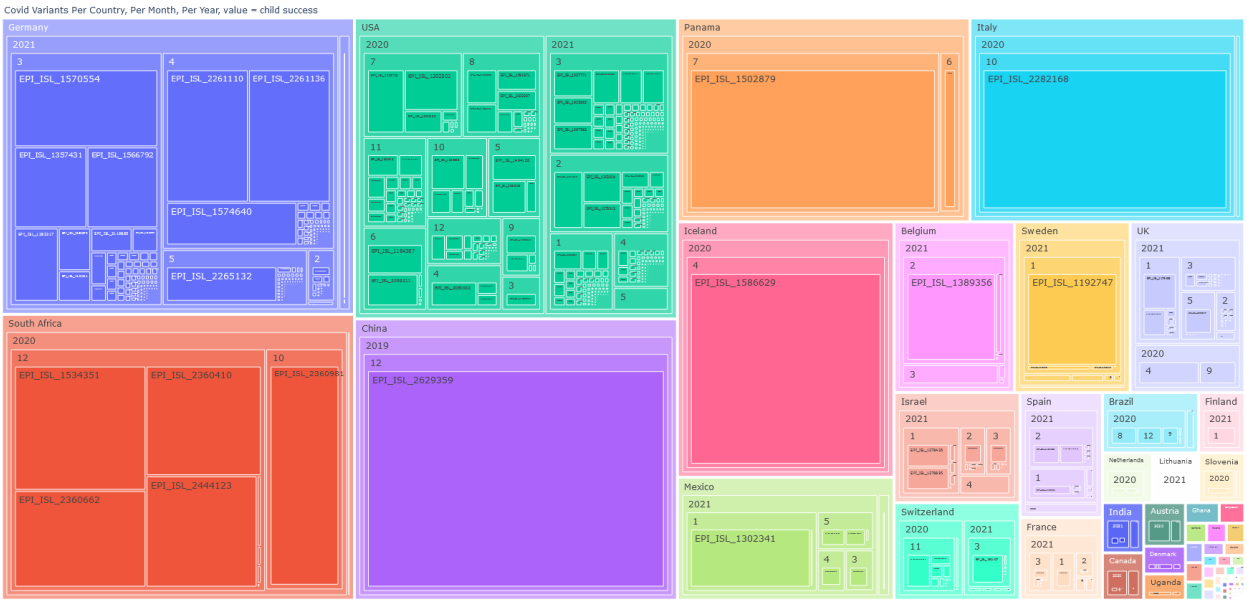


This Tree map sets each node's path based on year, accession where each box has a value of 1. The variations discovered in 2019 are included in this diagram, as well as all others but because only one variant discovered in 2019 had more than 30 recordings the 2019 box is too small to see. As you can see there are an incredible number of new variations in 2021 relative to 2020. The specific numbers are 2417 new variations in 2021 and 374 variations in 2020. Throughout all of 2019 and 2020, the world had a total of 84.3 million confirmed cases of Covid-19 (OurWorldInData). During the first 6 months of 2021 the world saw another 97.6 million confirmed cases of covid-19 (OurWorldInData). As a result it does not seem as though new covid variants linearly correlates between the number of new cases, as in reality the number of new cases during this time period only increased by around twice while the number of new variants increased by approx. 10 times. This treemap shows that at the very least, that the number of new covid mutations is exponentially increasing. This exponential trend is most likely related to the number of new cases as well as the number of covid variants active at any given time.

We also created 2 other treemaps to display similar information but with an added parameter being the country of origin. The first treemaps sets the value of each variation box to be 1, which is the same as in the previous treemap, but the second sets each variation to be set based on the success of each variation including children.



Through the above diagram we can observe a number of things. The first observation is that the U.S. has contributed the most globally to the total number of variants in both 2021 and 2020 which makes sense as we have the most total cases in the world during both of these periods (OurWorldInData). The second is that the previous observation that most variants were first recorded in 2021 holds true.



In the above treemap, similar to the previous map, each sequence is placed based on the country, year and month of first recording of each strain. In this treemap however each different mutation box size is based off of strain success as defined previously. This treemap suggests that even though the United States has produced the largest number of mutations, the new strains produced have only been somewhat more effective at reproduction than some strains first documented in other countries with less total strains. For example, a total of 47 new strains were documented originating from South Africa compared to the total of 1043 strains that originated from the USA. However, the total success of these strains is still comparable to the success of USA strains as South African strains and their children have been documented 904,063 times compared to the 870,846 times for USA strains. I think that this really points to how global this virus really is.

Conclusion

Given that the RBD, HR1, and HR2 segments began mutating around winter 2020-2021 I would conclude that it contributed to the rise in covid-19 cases. Given that the spike protein of the delta variant has an RBD mutation this seems likely. Since there are significant mutations in HR1, HR2, and the RBD subunit in late summer 2021 I would say that we are in for a surge for the next few months as well.

We can also conclude that the rate of mutations among covid variants is increasing rapidly most likely due to both the increase in covid cases but also in strain variations. As new strains are created, the rate of variations would also increase meaning that this trend is likely to continue into 2022 and is most likely exponential in nature. This conclusion can be reached by viewing any of the treemaps which all show that more variations have been created in the first half of 2021 than in both 2019 and 2020 combined.

In the future, it might be interesting if we could explore even more global data given that our current dataset was biased towards one country. We could also experiment more with other clustering algorithms or sequence comparison methods. Furthermore, experimenting with the new omicron variant would be very interesting. The omicron variant is possibly more deadly than the original virus and is far more infectious.

Bibliography

1. Huang, Y., Yang, C., Xu, Xf. *et al.* Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol Sin* **41,** 1141–1149 (2020). https://doi.org/10.1038/s41401-020-0485-4
2. H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, and M. Roser, "Coronavirus (COVID-19) cases - statistics and Research," *Our World in Data*, 05-Mar-2020. [Online]. Available: https://ourworldindata.org/covid-cases. [Accessed: 13-Dec-2021].
3. "Omicron variant: What you need to know," *Centers for Disease Control and Prevention*. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/variants/omicron-variant.html. [Accessed: 13-Dec-2021].
4. Lizhou Zhang, Cody B Jackson, Huihui Mou, Amrita Ojha1, et al. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity.
5. Shiyu He and Samuel W.K. Wong, 2021: Statistical challenges in the analysis of sequence and structure data for the COVID-19 spike protein.
6. Yuan Huang, Chan Yang, Xin-Feng Xu, Wei Xu, and Shu-wen Liu. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19.
7. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020;581:215–20.
8. "NCBI virus," *National Center for Biotechnology Information*. [Online]. Available: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Protein&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome%20coronavirus%202,%20taxid:2697049&Completeness_s=complete&ProtNames_ss=surface%20glycoprotein&SourceDB_s=RefSeq. [Accessed: 13-Dec-2021].
9. "Initiative," *GISAID*. [Online]. Available: https://www.gisaid.org/. [Accessed: 13-Dec-2021].
10. "Cumulative Confimed Covid-19 Cases". OurWorldInData. [online]. available:https://ourworldindata.org/grapher/cumulative-covid-cases-region. [Accessed: 12-Dec-2021].

Code Source:

https://github.com/PierceCappa/Covid-19-EDA-and-Visualization