Berkeley School of Information

W200 Project 2

Emerging SARS-CoV-2 Variant Mutation Analysis

Austin Jin, David Ristau, Matt Kinkley
4-15-2021

Table of Contents

1. Github Repository

Data and code for this project is available at our Github repository: https://github.com/UC-Berkeley-I-School/Project2_Jin_Ristau_Kinkley.git

2. Introduction

The COVID-19 pandemic has been raging for over a year now and will likely continue for the next months to come. Amid this ongoing crisis, there has been huge concerns with the new variants of SARS-CoV-2 that differentiate from the predominant variant already circulating among the general population. As the pandemic progresses with new variants of SARS-CoV-2 inducing public concerns, it is critical for the United States to sequence and analyze the new virus samples to draw a more cohesive picture of these newly discovered mutations. This work has been happening behind the scenes throughout the pandemic by epidemiologists across the world, but the story of how the virus has evolved throughout this pandemic has not been as easily available to the public. By parsing and analyzing raw data from credible sources, we intend to clarify the story of how this virus has evolved through the pandemic.

An initial review of the background of viral mutations suggests that mutations of viral genomes over time are not uncommon. This is such a common and generalized phenomenon that it is referred to as *Drakes Rule,* which states that the density of mutations per generation of an organism is inversely proportional to the genome size. Drakes rule applies to a broad category of organisms, including viruses [6]. SARS-CoV-2 is no different and has demonstrated genetic mutations as with all viruses starting from the very first observations. Based on lab samples collected from patients tied to the Wuhan market outbreaks, the SARS-CoV-2 virus has been shown to be mutating as early as when it was originally identified in the wet market outbreak in Wuhan China due to multiple strains of the SARS-CoV-2 Virus already circulating at that time [4]. In this paper we will explore various questions surrounding the current state of the various strains across the United States.

3. Research Questions

With this background knowledge in hand, our team sought to better understand the mutations that the SARS-CoV-2 virus has undergone through the course of this pandemic. To keep this research to a reasonable scope, the analysis was done with respect to the viral mutations within the United States. To develop a cohesive story on how these mutations have impacted the pandemic, we curated a series of guiding research questions:

1. What states have the most mutations over time?
2. What strain had the most mutations and is it uniform across all states?

3. What is the most predominant mutation in each state?
4. Is the number of mutations correlated with cases per capita?
5. Is any particular strain correlated with higher cases per capita?

4. Data Acquisition

To evaluate the research questions posed above, data on the total number of mutations, strains, and number of COVID-19 cases was required. Additionally, to evaluate the data in a proportional manner, data on the number of strains evaluated in the United States, and the population per state were also required.

The primary dataset leveraged in this analysis is sourced from Nextstrain, an open source project dedicated to pathogen genome data. Nextstrain is a highly reputable source that is funded in part by the National Institute of Health (NIH) [3]. This dataset provides a comprehensive breakdown of evaluated genomes of the SARS-CoV-2 virus as well identification of the specific strain, number of mutations, and other epidemiologically relevant information. The data is available to download from the Nextstrain website as a tab separated file with a total size of 1,127KB.

In addition to the information gathered from Nextstrain, a dataset from the CDC was used to provide further insight into the recent viral strains of concern. These variants have been seen in the news recently and we wanted to give them particular attention in our analysis. This data was gathered from the CDC website and contains information on the UK, South African, and Brazilian strains, and their frequency of detection in the US. This data is available for download on the CDC website as a .csv file and has a total size of 7 kb [7].

The COVID-19 case count data was sourced from an effort by the New York Times which publishes the total number of COVID-19 cases and deaths at the state, county, and national basis twice daily. This data is gathered from various local governments throughout the United States and made readily available by the New York Times in a convenient GitHub repository. In our analysis, we use the state level COVID-19 case data which is made available as a .csv file with a total size of 729 KB [5].

Data regarding the number of genomic sequences analyzed per state was gathered from the CDC. A goal of the CDC is to sequence 7,000 genomes per week to keep close tabs on the viral mutation of SARS-CoV-2 and this database tracks the achievement relative to that goal. The dataset consists of a list of states and the total number of genomes sequenced in each state. The website that houses this source data has changed throughout the duration of this project, but a similar file is now available for download from a tableau visualization on the CDC website for published COVID-19 sequences [1]. The original file used in this analysis is contained in the Github repository for this project. The file is available as a .csv file with a total size of 1KB.

The population of each state was sourced from the United Census Bureau through their ACS Demographics and housing estimates. This is an annual survey intended to update the population estimates throughout the United States. This data is available for download on the United States Census Bureau website as a .csv file with a total size of 162KB [2].

One final dataset was needed to aid in the construction of a summary dataset. The data from the various files contained full state names or abbreviations. This source file was acquired from the social security association website and the dataset is available to download as a .csv file with a total size of 12KB [9].

It should be noted that many of these data sources are update on a continuous basis, but for the sake of this analysis this data was collected between March 23rd and April 4th.

5. Data Exploration and Processing

Each dataset used in this project contained very clean data. The primary dataset sourced from Nextstrain is used in active analysis, which necessarily means much of the cleansing of raw input data is already completed. This is the case for each of the supplementary data sources as well. No observable discrepancies were found in the initial data plotting and exploration. Therefore, no data cleansing of each of the individual datasets was required. Upon combination of the data one cleansing activity that was required was to fill NaN values with zeros for an appropriate interpretation of the data, however this was primarily due to mathematical operations applied to the data not due to the presence of NaN values in the original datasets.

To start the exploration into the datasets that were gathered several initial graphs were produced. The number of cases and deaths from the beginning of year 2020 until March of 2021 were plotted in figures 5.1 and 5.2 below using data from the New York Times. The trend over time for this data shows a sharp upward trend in December 2020 in both cases and deaths.

Figure 5.1

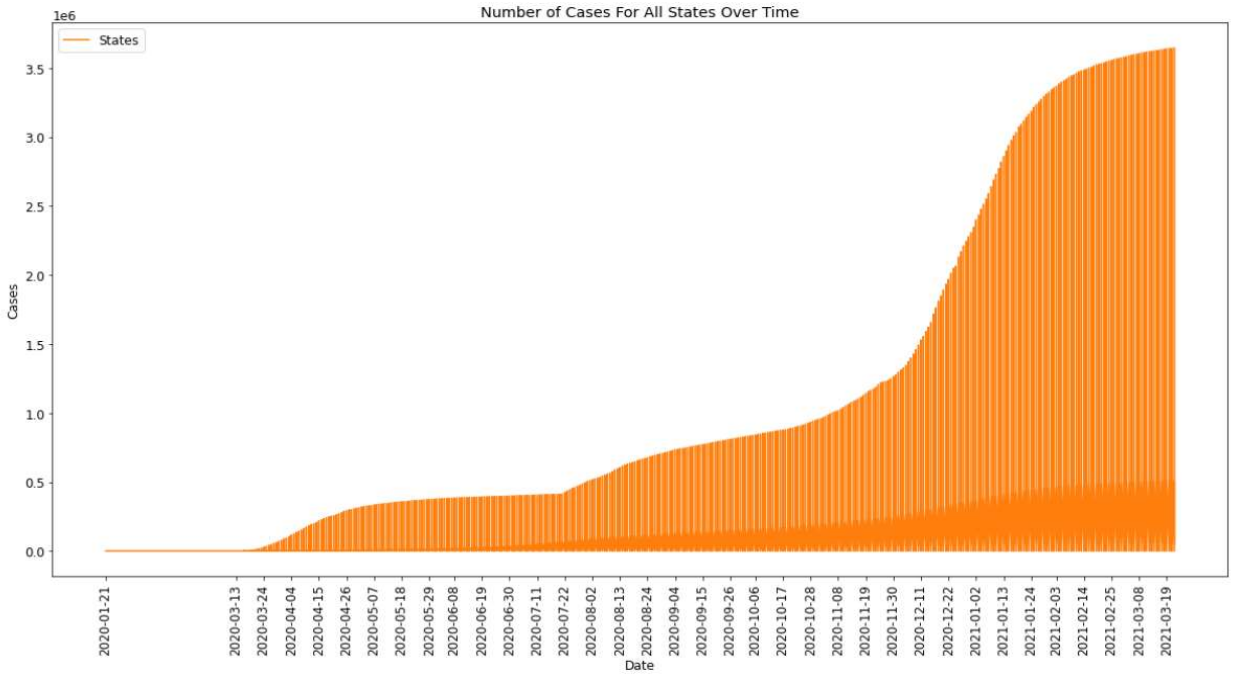
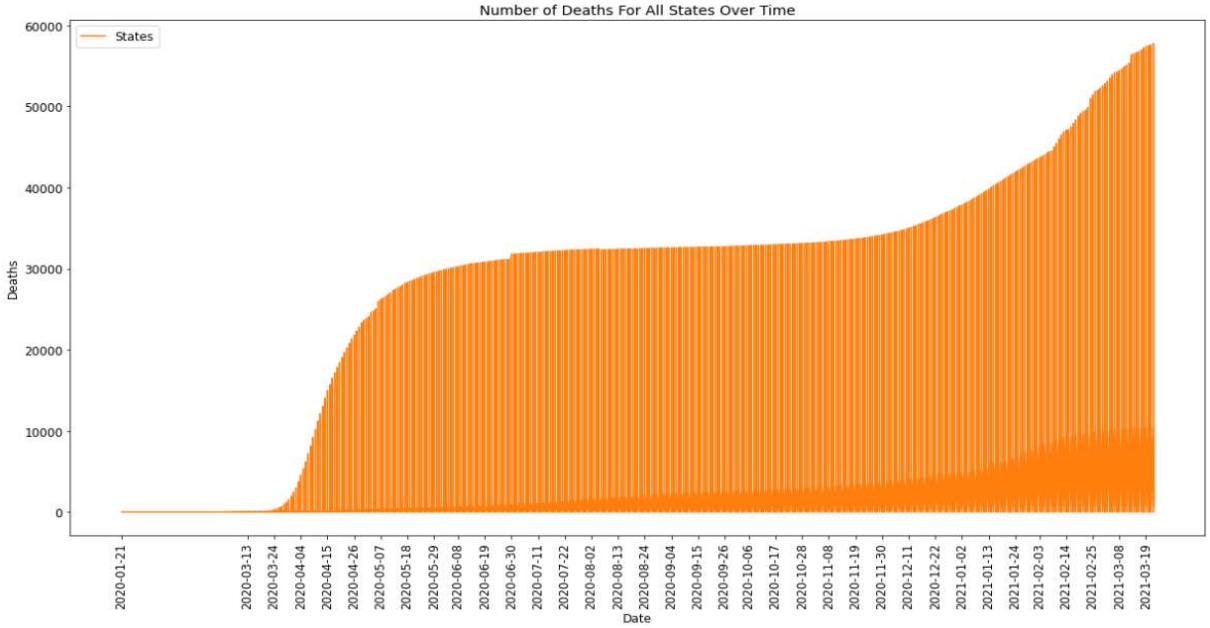


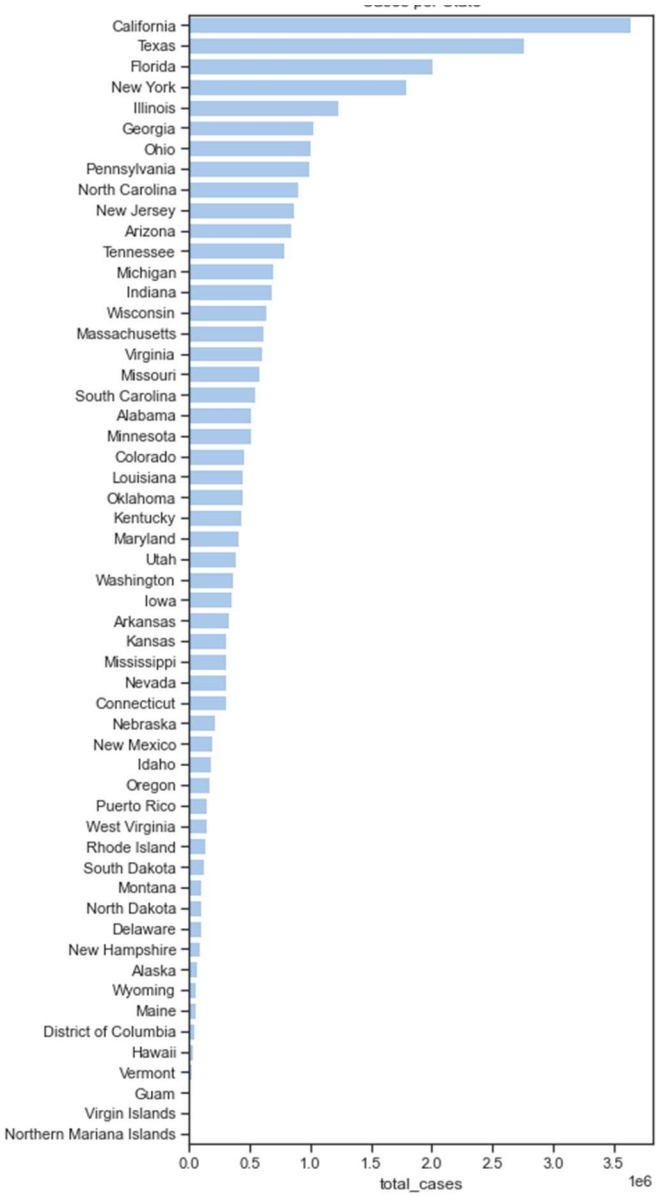
Figure 5.2



The total COVID-19 cases were plotted for each state are shown in figure 5.3. This data fits with the general story that has been shown in the news over the past year.

California, Texas, and Florida are states that are commonly referenced in the media and our data is reflecting a similar story indicating that those states have the highest total number of cases.

Figure 5.3



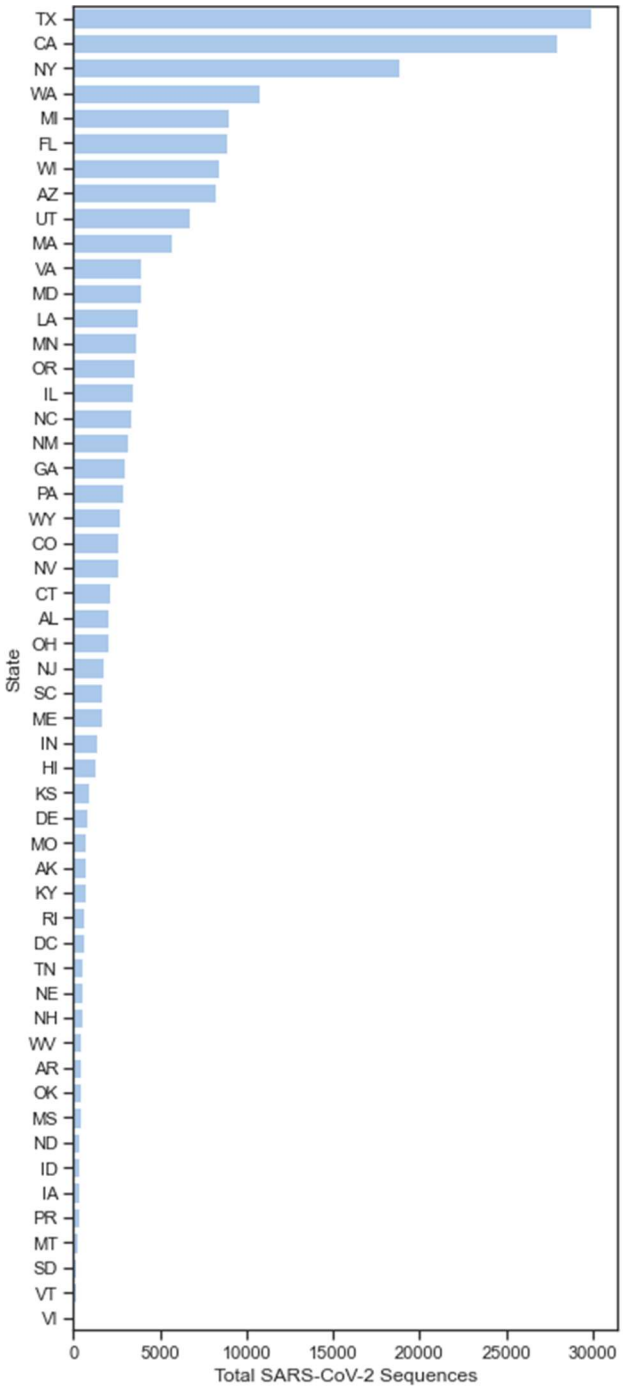
After browsing through the National Genomic Surveillance Dashboard from Centers for Disease Control and Prevention, we were able to initially obtain the numbers of sequences that have been evaluated in the United States and case data that provides information regarding the number of genomes that are sequenced in each state. Each state is plotted in terms of the total number of SARS-CoV-2 sequences that have been observed below in figure 5.4. This depiction shows that there is a large disparity in the number of genomic sequences that are evaluated in each state. This lack of uniformity across states is a potential limitation in the analysis of various strains on a state by state basis.

After the initial exploration of the datasets on and the primary task in the data wrangling portion of the analysis was to merge all datasets into a single data frame on which the analysis is based. This process was started by condensing the New York times dataset, which contained a rolling total COVID-19 cases and deaths in each state over time, into the total number of cases at the time the dataset was acquired. This process required only selecting the last datapoint for each state to populate a data frame with the most recent case and death numbers.

The NextStrain dataset was then grouped by states, and various features were extracted and merged with the state case count and death count data frame. This process entailed multiple iterations of grouping, sorting, and merging data frames to extract features for the analysis of each research question.

After the feature extraction from the Nextrain dataset, the additional information of state population, and the number of genomes sequenced in each state were required. The state population data was among the imported datasets and required a selection of the total population estimate for each state, and a merge with the summary dataset that contained all the Nextrain and COVID-19 features.

Figure 5.4



A unique challenge arose in merging the dataset with the total number of sequences due to the labels for each state only represented by a two-letter abbreviation. This required the use of another dataset containing each states full name and corresponding abbreviation. The abbreviations dataset was then merged with the total number of sequences which then allowed a common variable, the name of the state, to be used to combine this information with the summary data frame.

The summary dataset used for the primary analysis for this project is described in the codebook below in table 5.1.

Table 5.1

Field Name	**Description**
total cases	Represents the total number of COVID-19 cases per state.
total deaths	Represents the total number of COVID-19 related deaths per state.
19A	Each of the values in this section represent the number of each clade observed in each state.
19B	
20A	
20B	
20C	
20D	
20E (EU1)	
20G	
20H/501Y.V2	
20I/501Y.V1	
1	Each value in this column represent a count of the number of clades identified to have each number of s1 mutations.
2	
3	
4	
5	
6	
7	
8	
total_s1_mutations	Represents the total number of observed spike protein mutations in each state.
total_pop_2019	Represents the total population in each state sourced from the census bureau.
cases_per_cap	Represents the cases per capita in each state.
Total SARS-CoV-2 Sequences	Represents the total number of genomic sequences reported by each state sourced from the CDC.
s1 mutations per observation	Represents the number of s1 mutations per observation as a measure of the rate of mutations in each state.

6. Analysis and Discussion

After gaining an initial understanding of the spread of SARS CoV-2 and developing a summary dataset, we started our analysis of the evolution of the viral strains throughout the COVID-19 pandemic. Prior to this analysis a set of basic vocabulary must be established to enable a reader to develop useful insights from this analysis. To provide this common vocabulary and establish some background information, we provide the following summary of genetic mutation and virus terminology:

When a virus mutates it means that the basic genetic makeup of the virus remains the same, with small changes to subcomponent genetic makeup of the virus. These mutations are referred to as either strains or clades. A strain and a clade are, at their core, the same thing: a genetic mutation of the virus. A mutation is referred to as a clade when it has no impact on how the virus interacts with its host. When a mutation does have an impact on how the virus interacts with its host it is referred to as a strain [10]. Therefore, each strain may also be referred to as a clade, but not all clades can be called a strain. A mutation that is of primary interest in the SARS-CoV-2 virus is called an S1 mutation. ‘S’ refers to the distinctive spike protein that gives the corona virus its name. This protein is the part of the virus that is responsible for infecting host cells thus causing the infection. An S1 mutation is a mutation of the subcomponent in the spike protein which is responsible for binding with target cells [8].

With this background information and common vocabulary established, we proceeded to answer each of our research questions.

1. What states have the most mutations over time?

To answer this question, we first define mutations to mean S1 mutations. This clarification is made because S1 mutations have most significant impact on how infectious the virus is, making this type of mutation of particular interest. Additionally, S1 mutations is reported in the Nextsrain dataset and can be grouped by state to identify which states have the most mutations. Each genome that is evaluated has a corresponding number of mutations that are recorded. The distribution of mutations for each genome is shown below in figure 6.1. The most frequently reported number of mutations on a given viral genome is 1, but in some instances as many as 8 mutations were found, though this is rare.

Data representing the frequency of S1 mutations was gathered and presented on a state-by-state basis. To compare each state on an equivalent basis, the total number of S1 mutations that were observed was divided by the number of genomes that were sequenced in each state. Figure 6.2 demonstrates that South Dakota has a much higher rate of observed sequences, followed by Iowa and Rhode Island, after which the rate of observed mutations reduces quickly.

Figure 6.1

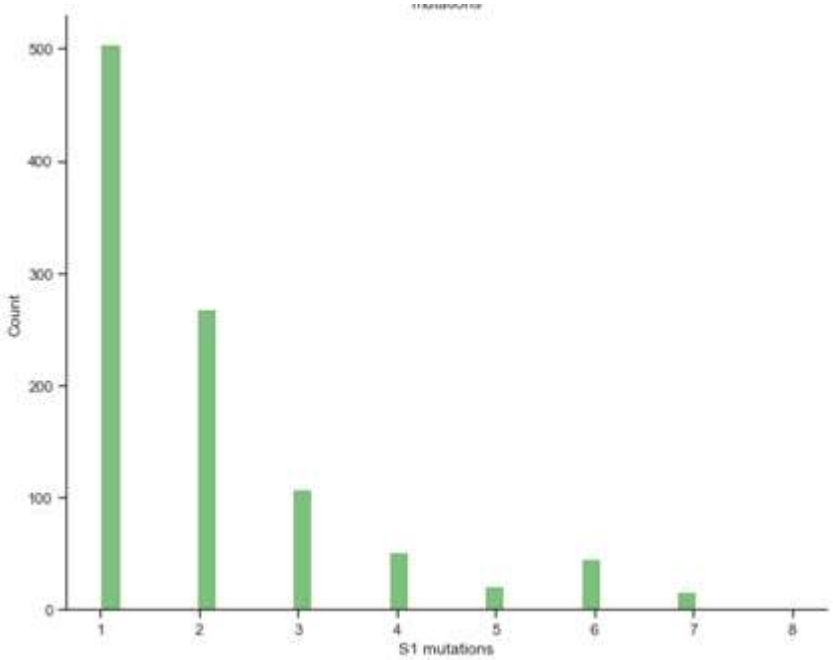
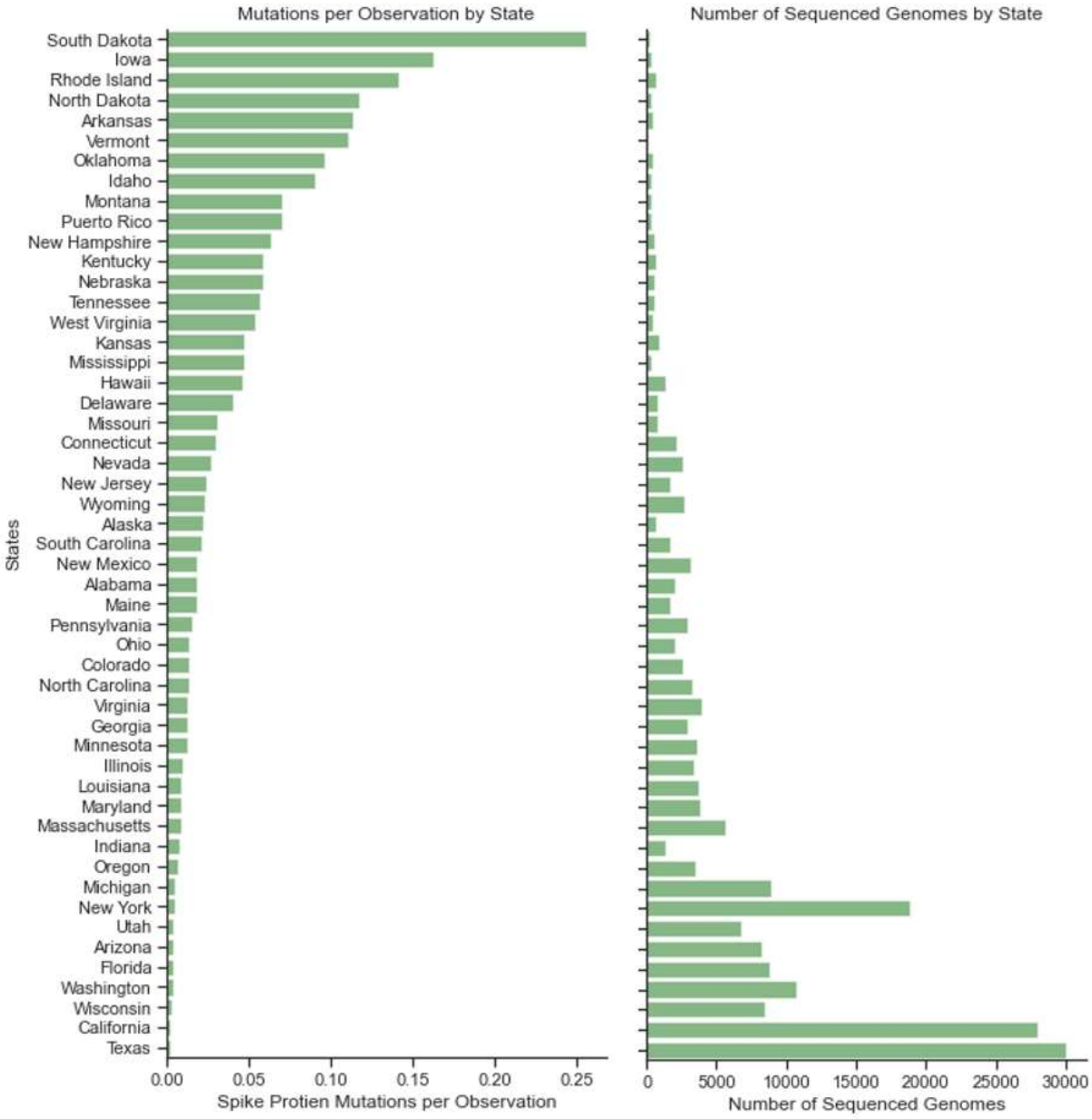


Figure 6.2



When the number of genomes sequenced in each state are plotted next to the spike protein mutations per observation, it demonstrates that the three states with the highest rate of observed mutations appear to have the fewest number of genomes sequenced. This is concerning as it indicates that because of the low sample size this may only be due to statistical noise. If the states of consideration are limited to only those states that have sequenced over 5000 genomes, we can assume that the true number of mutations has converged, or at least come closer to doing so, based on the law of large numbers. In the more limited grouping of states with over 5000 genomes sequenced we see that Massachusetts appears to have an unusually high rate of observed S1 mutations. This conclusion should be taken tentatively due to the inverse relationship between

sequences and spike protein mutations. To gain a more accurate interpretation of this phenomena, more uniform testing would be required among the states, particularly those who have less than 5000 genomes sequenced.

2. What strain had the most mutations and is it uniform across all the states?

Figure 6.3 below shows that the clade with the highest number of mutations across the country is 20C followed closely by 20G. Notably, clade 20I/501Y.V1 has the third highest number of mutations. 20I/501Y.V1 represents the UK variant which has been circulating for far less time in the United States than either 20C or 20G. This indicates that while 20C has the highest number of observed mutations at the time of this analysis, clade 20I/501Y.V1 is mutating more aggressively.

Figure 6.4 below shows a heatmap of S1 mutations per clade per state. This heatmap shows that the clades with the most observed mutations are consistent with figure 6.3. In general, most states have observed the most mutations with the 20C clade, with some states observing 20G, 20A and 20I/501Y.V1.

Figure 6.3

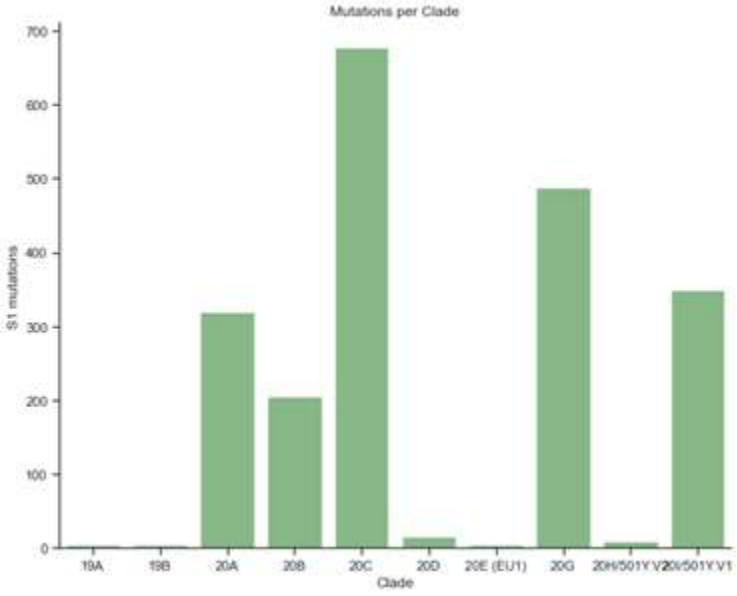
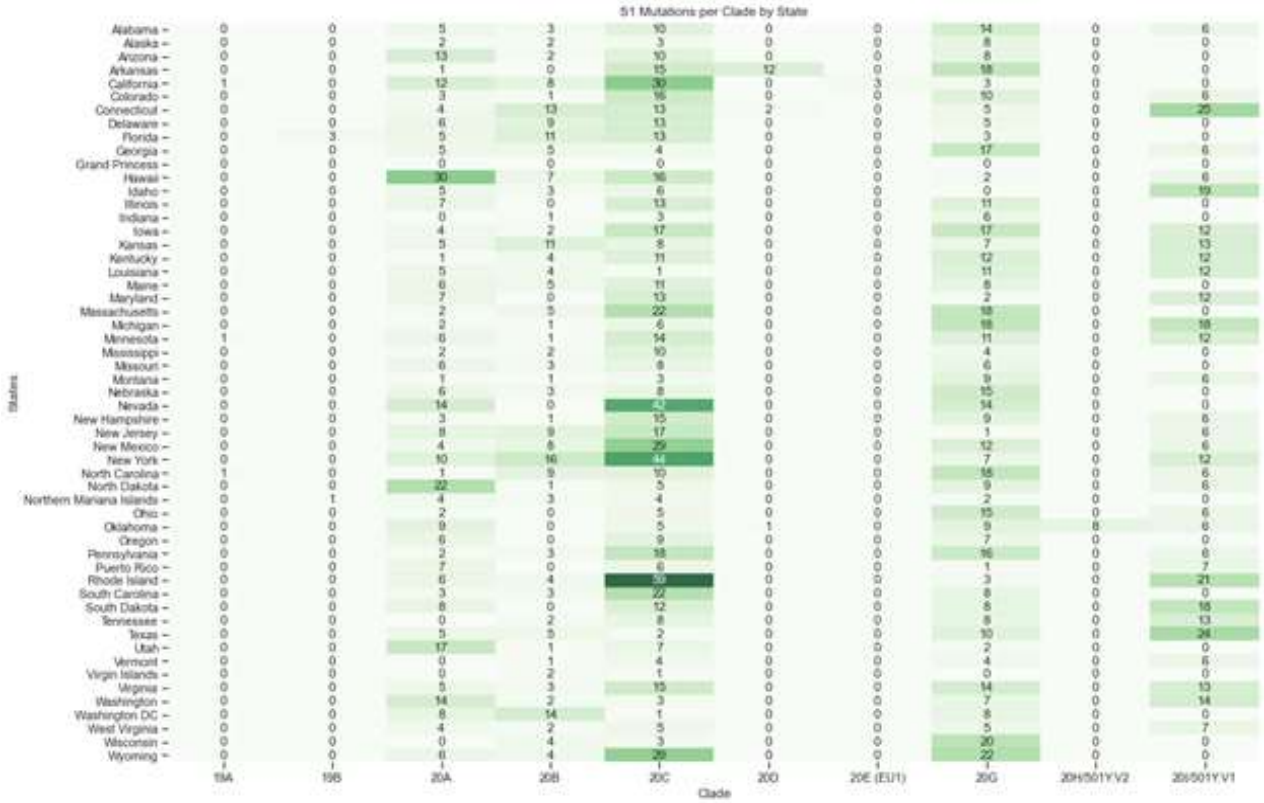


Figure 6.4

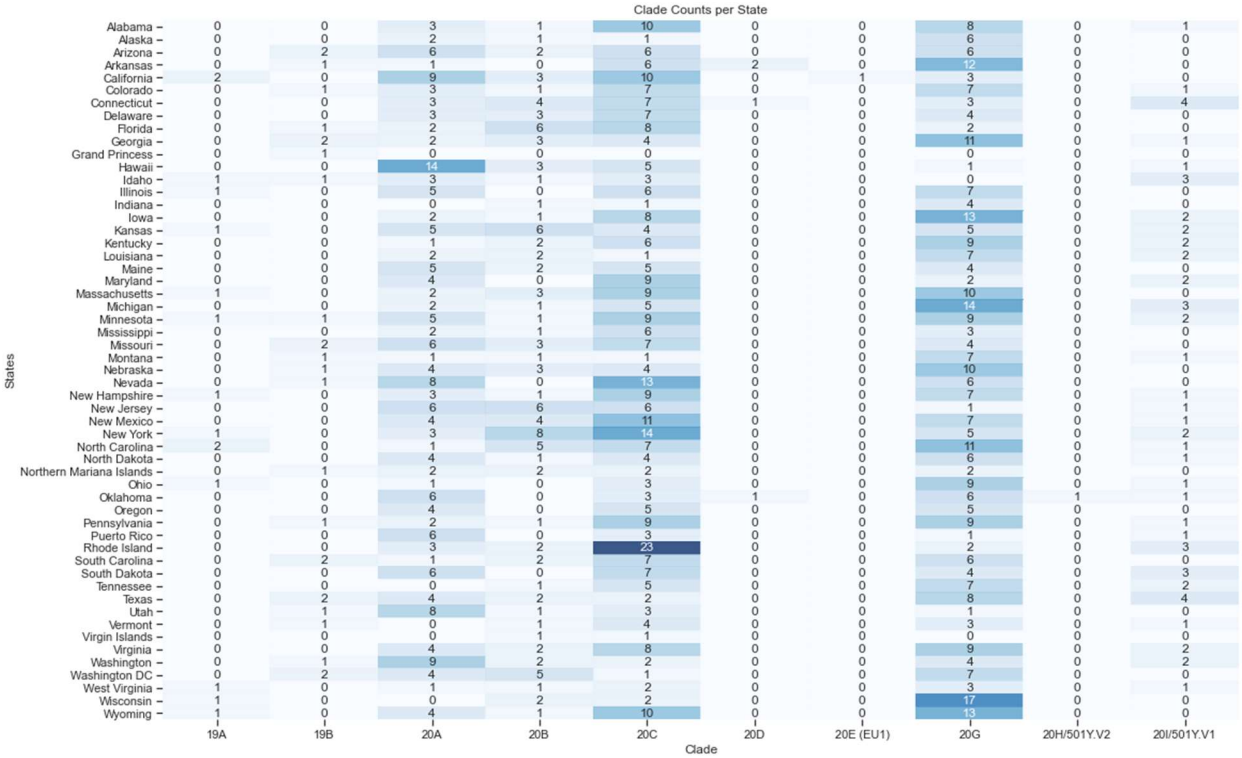


This analysis demonstrates that at current, the 20C clade has the greatest number of mutations across most of the country, however 20I/501Y.V1 is mutating rapidly and may soon overtake 2C.

3. What is the most predominant mutation in each state?

To answer this question a heatmap of the count of each clade across all states was generated. This heatmap can be seen below in figure 6.5. In general, we see that the most prominent clades are 20C and 20G and this is relatively uniform across all states. Some notable exceptions to this are Alaska and Washington which appear to have the highest number of the 20A clade.

Figure 6.5



To further diversify our analysis the following three variant types: B.1.1.7, B.1.351, and P.1 were analyzed. These strains represent ‘variants of concern’ that the CDC is monitoring closely and represent variants from the UK, South Africa, and Brazil respectively [7]. We analyzed the number of reported cases of these three variants using a separate dataset: the CDC’s “Cases of Variants in the United States”. In order to assess which top 10 states held the highest number of each strain, the number of observations of each strain are plotted in figure 6.6, 6.7, and 6.8. Our initial hypothesis was that if a certain state had a relatively high number of a certain variant, that state would also have a high number of other variants of concern.

Figure 6.6

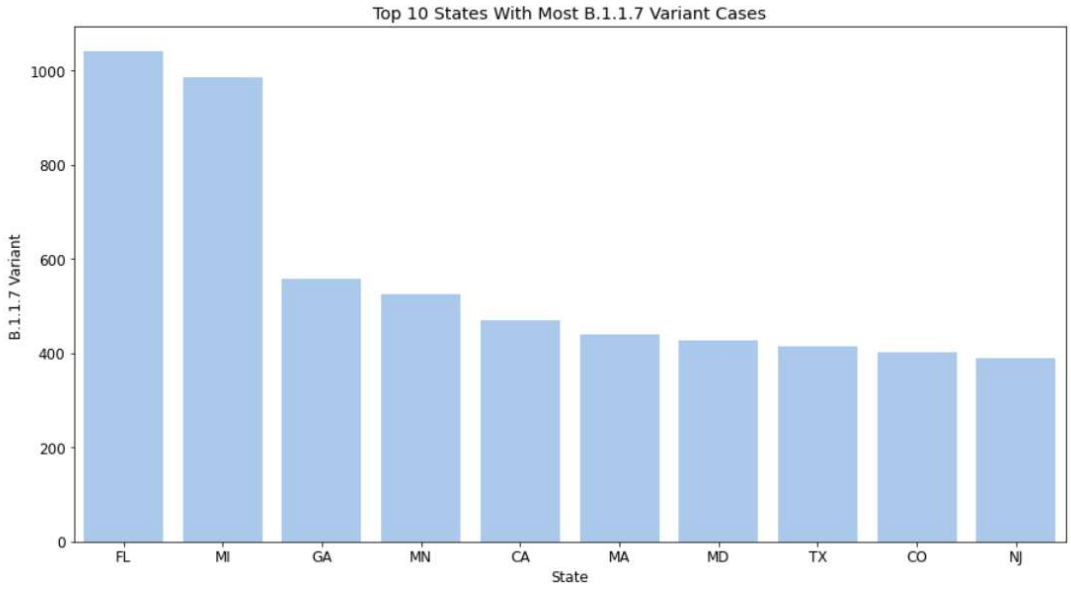


Figure 6.7

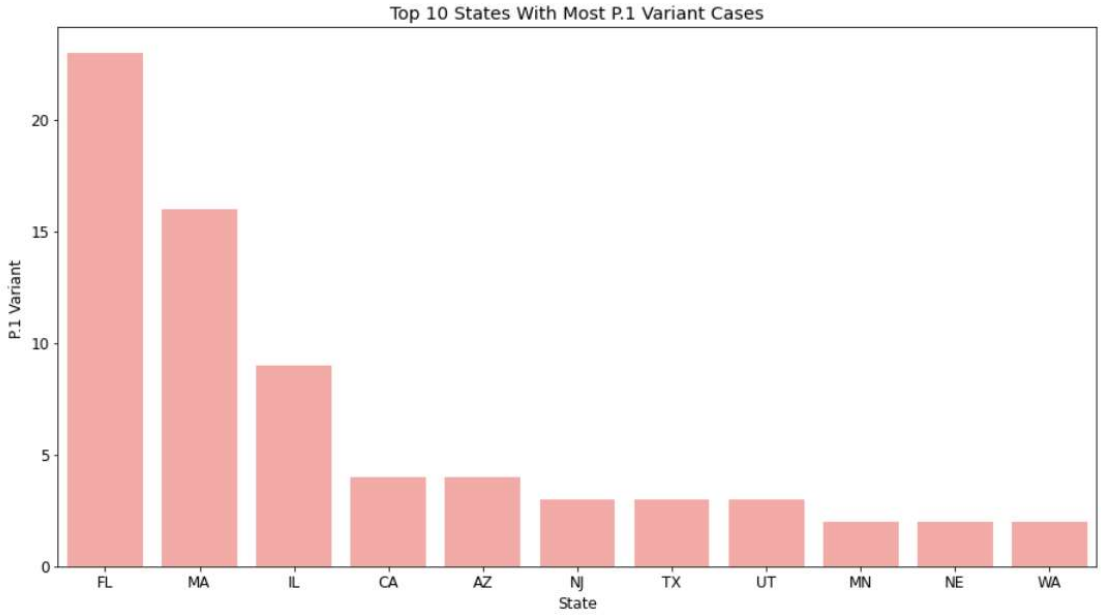
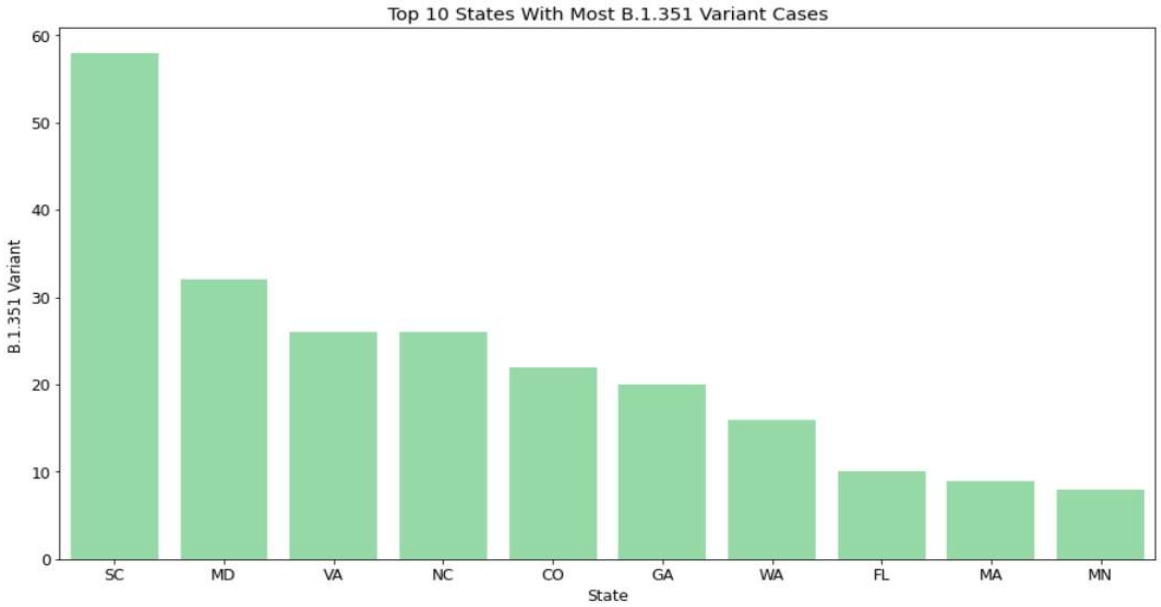
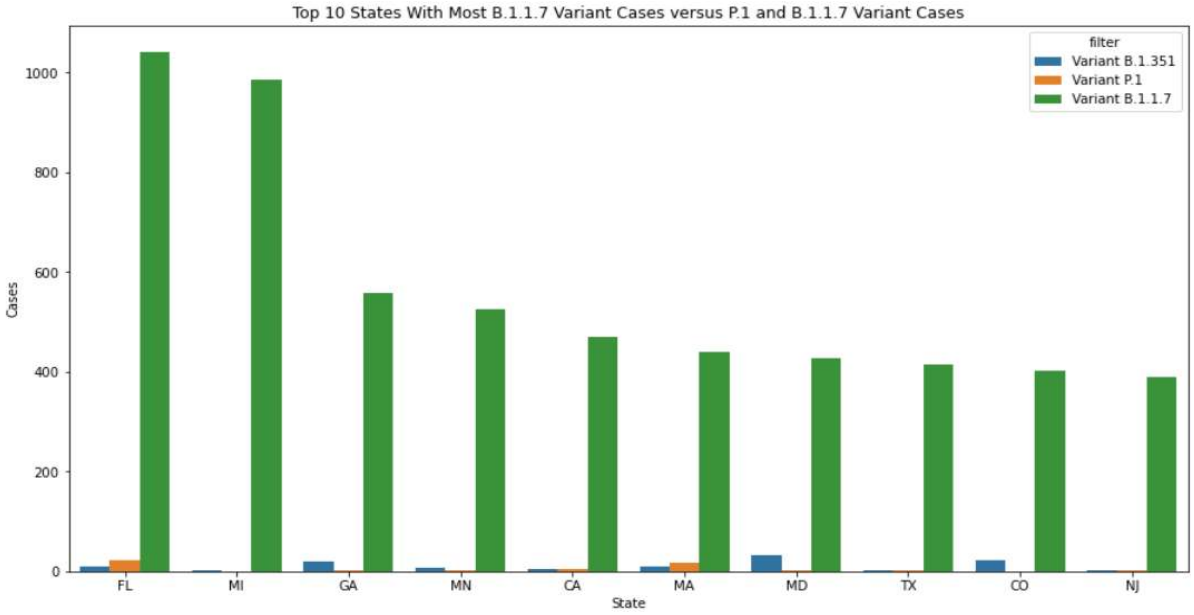


Figure 6.8



As you can see from the figures above, a high occurrence of one variant does not necessarily mean there is a high occurrence of another variant in each state. The UK variant, B.1.1.7 had a particularly high case count in Florida, as did the P1 variant but Florida is much lower on the list for the South African variant. This can be further demonstrated when each of the three variants are plotted together in figure 6.9. In this visual the sense of scale of each variant is better demonstrated.

Figure 6.9



As you can see from this analysis, Variant B.1.1.7 held the highest number of cases by far compared to the other two variants of P.1 and B.1.351 in the states of FL, MI, GA, MN, CA, MA, MD, TX, CO, and NJ. However, a higher rate of observations of one variant did not correspond directly to an increase in the number of cases of another variant.

After analyzing the phylogeny root of each variant on the Nextstrain interactive subsampling dashboard we were able to depict that the reason may be because each of the three variants aren't biologically related and originate from different continents. B.1.1.7 is from the United Kingdom, B.1.351 is from South Africa, and P.1 is from Brazil [3]. This could indicate that travelers from various parts of the world brought these variants to the united states, but each one likely had an independent origin.

Through this analysis we have evidence say that the most predominant strain right now in each state are the 20C and 20G clades, however the B.1.1.7 variant is spreading quickly and may soon become the predominant variant.

4. Are mutations correlated with cases per capita?

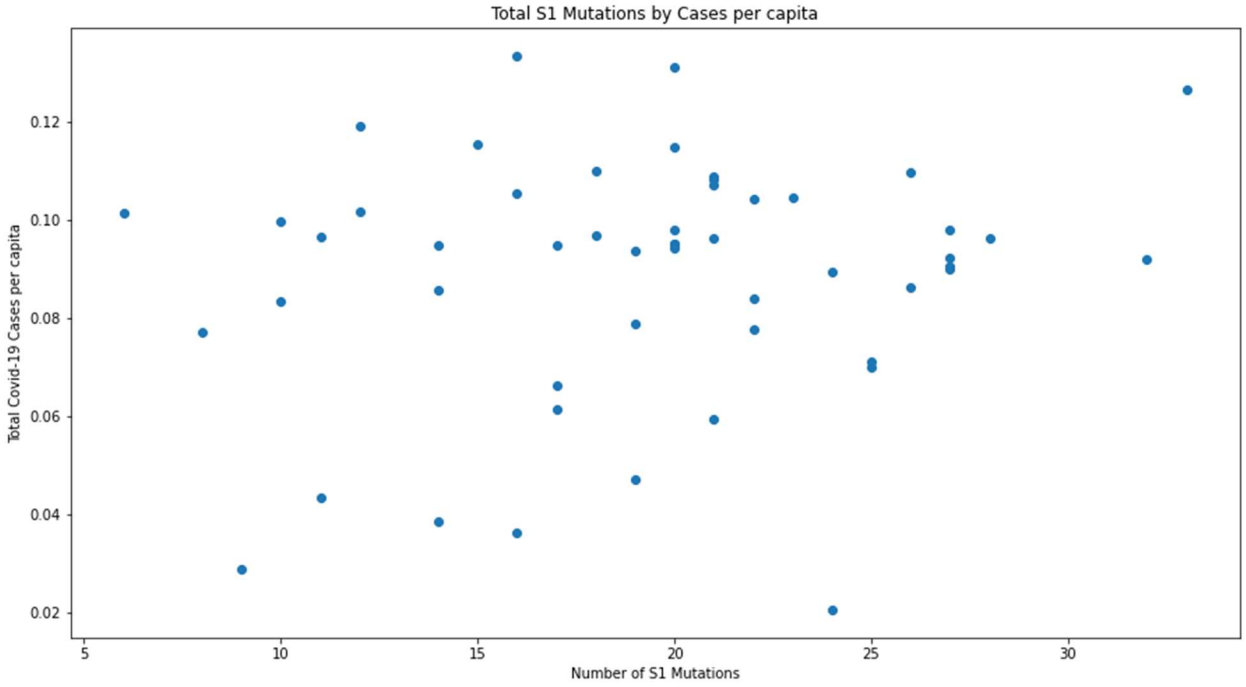
An example of the data frame used to investigate the following analyses is shown below:

Table 6.1

state	total_cases	total_deaths	total_pop_2019	cases_per_cap	19A	19B	20A	20B	20C	20D	20E (EU1)	20G	num_clades	total_s1_mutations
California	3645558	57744	39512223.0	0.092264	2.0	0.0	9.0	3.0	10.0	0.0	1.0	3.0	28	27.0
Texas	2758353	47616	28995881.0	0.095129	0.0	2.0	4.0	2.0	2.0	0.0	0.0	8.0	22	20.0
Florida	2011203	32778	21477737.0	0.093641	0.0	1.0	2.0	6.0	8.0	0.0	0.0	2.0	19	19.0
New York	1787934	48959	19453561.0	0.091908	1.0	0.0	3.0	8.0	14.0	0.0	0.0	5.0	33	32.0
Pennsylvania	993622	24839	12801989.0	0.077615	0.0	1.0	2.0	1.0	9.0	0.0	0.0	9.0	23	22.0
Illinois	1226595	23379	12671821.0	0.096797	1.0	0.0	5.0	0.0	6.0	0.0	0.0	7.0	19	18.0
Ohio	1001195	18340	11689100.0	0.085652	1.0	0.0	1.0	0.0	3.0	0.0	0.0	9.0	15	14.0
Georgia	1021299	17990	10617423.0	0.096191	0.0	2.0	2.0	3.0	4.0	0.0	0.0	11.0	23	21.0
North Carolina	903311	11865	10488084.0	0.086127	2.0	0.0	1.0	5.0	7.0	0.0	0.0	11.0	27	26.0
Michigan	696800	16904	9986857.0	0.069772	0.0	0.0	2.0	1.0	5.0	0.0	0.0	14.0	25	25.0

First focusing on mutations, we took the S1 mutations reported for each state and plotted it against their per capita cases as shown in figure 6.10 below.

figure 6.10

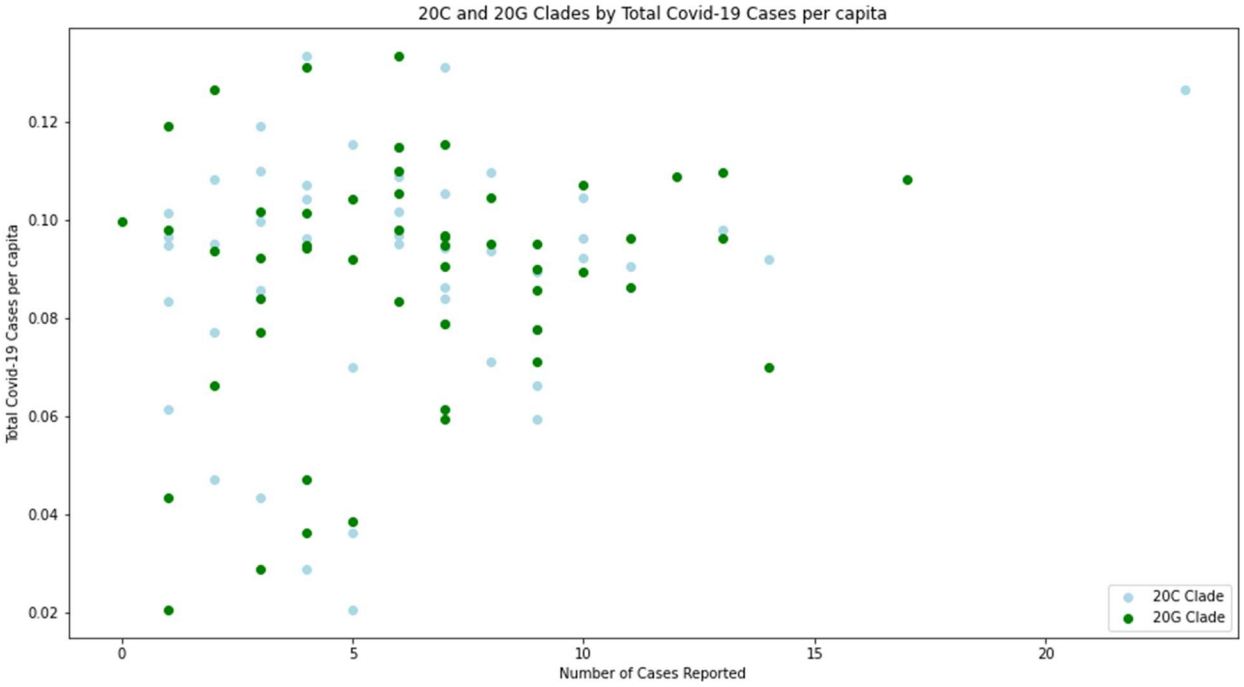


We chose a scatter plot to better visualize the relationship between the number of mutations and per capita cases. Each dot in figure 6.10 represents a state, with their number of S1 mutations reported on the x-axis, and their total cases per capita on the y-axis. As we can see, there is not a strong relationship between these two variables. There are quite a few states which have a high normalized count of COVID-19 cases, but do not have as many S1 mutations reported relative to other states. Based on the data represented in this scatter plot we cannot draw any conclusions with confidence with the relationship between the number of mutations and the number of cases per capita.

5. Is any particular strain correlated with higher cases per capita in a given state?

To answer this question we ran a similar analysis as that from question 4. However, this time focusing on 2 of the top clades to see if individual strains might be correlated with higher total case counts. Taking only the 20G and 20C clades as they are most prominent, we plotted these counts for each state against the total cases per capita in figure 6.11.

Figure 6.11



Again, we do not see a strong correlation, however it should be noted that in general the more cases reported for these strains, the more likely that state will have a high case rate per capita. We do not see the opposite effect though; there are still many states with low cases for these strains yet high total cases per capita. Due to the relatively small number of cases reported for specific clades (possibly due to testing availability), it is hard to draw a definitive conclusion. However, with more sample data collected on cases for particular strains, we expect to see a stronger positive correlation.

7. Conclusion

In answering each of our initial research questions we have demonstrated a far more complex story of continuous mutation than would be suggested by the level of coverage in common media sources.

The answer to our first two questions tell a story that there is a large rate of mutation in the spike protein, which is particularly prominent in the 20C and 20G clades. Additionally, we see that the new UK variant (strain B.1.1.7/ Clade 20I/501Y.V1) is mutating very quickly and may soon catch up to 20C and 20G. Our analysis of question three demonstrates that at current, the 20C and 20G clades are the most common throughout the United States. Although, we also see indications that the UK variant is on track to quickly overtake both of these clades. Our answers to questions four and five are inconclusive but suggest potential relationships may be present, but not identifiable in the data that was used.

We believe that our report marks step forward in more effectively communicating the complex and critical research that is currently being performed by epidemiologists across the country. This information is more important now as we continue to roll out vaccines across the country. The race is now between genetic mutation and our ability to achieve herd immunity. An informed public will be better able to understand the need for continuous vigilance against this virus and may be more receptive to future vaccination efforts.

8. References

1. CDC COVID Data Tracker. (n.d.). Retrieved March 25, 2021, from https://covid.cdc.gov/covid-data-tracker/#published-covid-sequences
2. Data.census.gov. (n.d.). Retrieved April 1, 2021, from https://data.census.gov/cedsci/table?q=ACS&g=0100000US.04000.001&tid=ACSDP1Y2019.DP05&moe=false&hidePreview=true
3. Nextstrain. (2021, April 8). Retrieved April 15, 2021, from https://nextstrain.org/
4. Novella, S. (2021, March 31). The origins of sars-cov-2. Retrieved April 13, 2021, from https://sciencebasedmedicine.org/the-origins-of-sars-cov-2/
5. Nytimes. (n.d.). Nytimes/covid-19-data. Retrieved March 25, 2021, from https://github.com/nytimes/covid-19-data/commits?author=nyt-covid-19-bot
6. Sanjuán, R., Nebot, M., Chirico, N., Mansky, L., & Belshaw, R. (2016). Viral mutation rates. *Virus Evolution: Current Research and Future Directions,* 1-28. doi:10.21775/9781910190234.01
7. Sars-cov-2 variants of concern. (n.d.). Retrieved March 28, 2021, from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html#Concern
8. Sasaki, M., Uemura, K., Sato, A., Toba, S., Sanaki, T., Maenaka, K., . . . Sawa, H. (2021). Sars-cov-2 variants with mutations at the s1/s2 cleavage site are generated in vitro during propagation in tmprss2-deficient cells. *PLOS Pathogens, 17*(1). doi:10.1371/journal.ppat.1009233
9. Social security. (n.d.). Retrieved March 25, 2021, from https://www.ssa.gov/international/coc-docs/states.html
10. Zoppi, L. (2021, March 23). Viral clades of sars-cov-2. Retrieved April 1, 2021, from https://www.news-medical.net/health/Viral-Clades-of-SARS-CoV-2.aspx