**Lab 5 - Evolutionary analysis of COVID-19 genomes**

Author: Yanni Guo

Date: December 2021

**Introduction:**

This lab report will discuss how researchers use tools to sequence SARS-CoV2 genomes to find variants in the genomes and observe whether the variants changed over time. The SARS-CoV2 genomes were collected from random samples of patients in England from November and December 2020. This lab consisted of five different parts to analyze the variants of SARS-CoV2 genomes. We used the National Center for Biotechnological Information (NCBI) tool to find mutations and individual mutations from the multiple sequence alignment of the SARS-CoV2 genome. We created functions to create bar graphs, heat maps, and a tree to study the change in the variants over time.

**Methods:**

We used the NCBI tool to study the multiple alignment sequences of the SARS-CoV2 genome for parts one and two of the lab. We created three different graphs in Python to answer questions for parts three, four, and five of the lab. The process of using the NCBI tool and creating the graphs is discussed in this method section. All the procedures in using the sequence alignment tool and constructing the graphs, parameters, variables, and algorithms are explained to answer the lab’s exercise questions and overall goals.

For part one of the lab, “Finding mutations”, we used the NCBI sequence alignment tool to answer exercise questions one and two (NCBI, “NCBI Multiple Sequence Alignment Viewer 1.21.0.”). We uploaded the November 3 fasta file into the NCBI tool to view the multiple alignments of the SARS-CoV2 genome. The NCBI tool showed 100 rows of alignments of the fasta file. To answer question one, we zoomed in on the red and dark grey vertical regions in the multiple alignments. Then we clicked on the red and dark gray regions of the alignments to obtain more information of specific nucleotide positions in those regions, such as the base, matches, mismatches, gaps, unaligned, etc. To answer question two, we zoomed out to get a general overview of the multiple alignments to find any patterns.

For part two of the lab, “Profiling individual mutations”, we continued to use the NCBI tool and uploaded the December 8 fasta file to answer questions three through six. We zoomed in on the multiple alignments, clicked on bases at certain positions, and compared certain base positions between the November 3 and December 8 fasta files of the SARS-CoV2 genome. We were able to determine original bases, mutations, nucleotide sequences, and deletions at specific alignment positions.

For the “Comparing variant prevalence over time” part of the lab, we first looked carefully into the alignment results from the NCBI tool to find the pattern and position for each variant. We found that within the results from the samples collected in all five days, the positions where each variant shows up are the same across days except for the ones on December 5th (which is always a bit off behind). Then we use the alignment fasta files we download from the NCBI tool as inputs for our program, to count the number of reads that satisfy all three variants on each date. Then use the ratio between count and the number of samples we have each day to plot the proportion of B.1.1.7 mutation for each date’s samples.

For part four of the lab, “Visualizing variant changes with heat maps”, we used the python script titled “heatmap.py” provided on Notebowl to construct the heatmaps. In the code, we put in the correct inputs of the “HCOV19-ENGLAND-xxxx20-D.pim” files for each corresponding date to create the heatmaps. To answer question eight, we observed the December 8, 2020 heatmap’s interesting features, determined what the different colored-value features represent, and compared the December 8 heatmap with the rest of the November and December heatmaps.

For the “Visualizing variant changes with evolutionary trees” part, we wrote an “exercise 9” function and 8 helper functions to perform the neighbor-joining algorithm with the “-D.fasta” files as input. The output result for these functions is the trees (one tree for a day’s samples) presenting the distances in between all the samples we have for each day, showing how close/similar the samples are to each other, and what are the outlier samples that might be the variant changes.

**Results:**

For all five parts of the lab, we answered exercise questions 1 to 9 to understand the variants and change in variants over time in the multiple alignments of the SARS-CoV2 genome by using the NCBI tool and creating three different graphs. This part of the report displays the results for the lab questions and interprets the results.

In part one of the lab, we addressed questions one and two. To answer question 1, the red vertical bars in the multiple alignments represent a single base position where a mutation occurred. The light gray regions represent the matched bases. The dark gray regions represent the uncertain bases at specific positions in the multiple alignments due to mismatches. The red vertical bars only had one base match for certain positions, and the single base mutations caused mismatches within the multiple alignments. The dark gray regions have very low numbers of matches and a high number of mismatches, which meant there were multiple base options for certain positions in those regions. When comparing the November 3 and December 8 fasta files, red and dark gray regions appeared more frequently as time increased. This may indicate that more variants in the SARS-CoV2 genome increased over time. Question 2: yes, we found some regions of the genome that are interesting in terms of viral variation. For example, at alignment positions 21,613 and 22,444, there were consecutive red regions in the same position at different sequences. The stacks of red regions at the same alignment positions formed red vertical lines in the zoomed-out view of the multiple alignments, which questions the base mismatches at those certain positions. Since the mutations appeared in the same column, the data suggested that the mutations were not random.

In part two of the lab, we answered questions three through six. Question 3: table 1 showed the base and mutation information at alignment position 23,063 to answer question 3. The number of November 3 and December 8 genomes with original and mutated nucleotides were out of 100 genomes.

|  |  |
| --- | --- |
| Presumed original nucleotide | Adenine |
| Number of November 3 genomes with original nucleotide | 99 genomes |
| Mutation of the original nucleotide | Thymine |
| Number of December 8 genomes with mutation | 46 genomes |

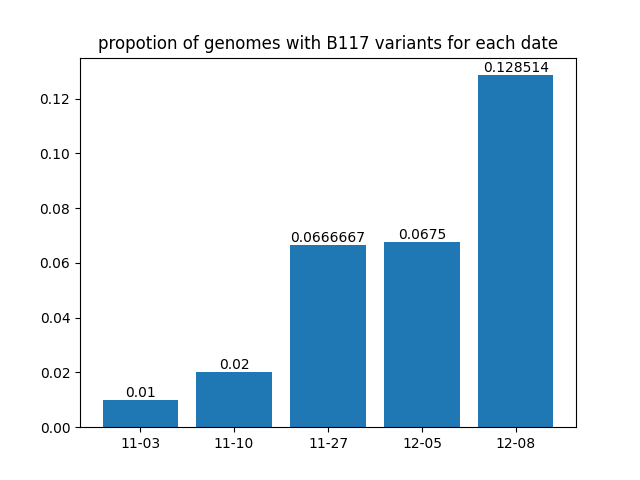
**Table 1:** Table shows the nucleotide information and mutation at alignment position 23,063.

Question 4: “TACATG” was the nucleotide sequence that was deleted from both the November 3 and December 8 genomes at positions 21765 through 21770. Question 5: the mutations of the six nucleotides from question 4 were most likely deletions. Because based on the gaps in the November 3 alignment, the mutations were missing from the first genome alignment. Further, there were only six genome alignments that had the “TACATG” gap in the November 3 alignment and fifty-three genome alignments that had the “TACATG” gap in the December 8 alignment. Which meant more of the “TACATG” nucleotide sequence got deleted from the multiple alignments as time passed. Question 6: table 2 showed the base and base mutation information at alignment position 23,604, which are the response to question 6. The number of November 3 and December 8 genomes with original and mutated nucleotides were out of 100 rows of genomes alignment.

|  |  |
| --- | --- |
| Original nucleotide | Cytosine |
| Number of November 3 genomes | 99 genomes |
| Mutation at the position | Adenine |
| Number of December 8 genomes with mutation | 45 genomes |

**Table 2:** Table shows the base and mutation information that responses to question 6 of the lab.

For part three of the lab, we created a bar graph. From the bar plot below (Figure 1), we can clearly see that as time passes, the proportion of B.1.1.7 variants for each day is increasing, and the rate it increases also rises. Since we have roughly 100 samples for each day, the number of B.1.1.7 variants collected for a day increased only by 1 for the first 10 days in November. But the number of B.1.1.7 variants per day increased by roughly 6 values within three days between December 5th and 8th. This means the B.1.1.7 variants probably spread much more rapidly as time passes.

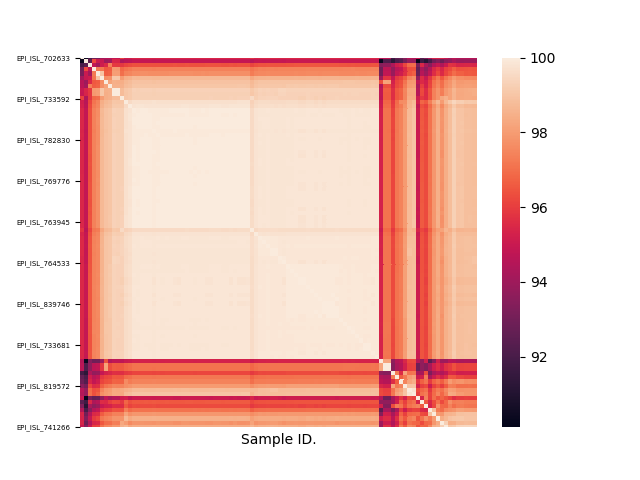


**Figure 1:** Bar graph that answers question seven. The X-axis represents the sampling dates of the genomes. The Y-axis represents the proportion of genomes that have all three of the B.1.1.7 variants.

For part four of the lab, *Visualizing variant changes with heat maps,* we created heatmaps to answer exercise question eight. Interesting features of the December 8th, 2020 heatmap figure (Figure 2) include the variation of colors for certain sample ID pairs. For samples in the file that were placed between the ones with IDs as EPI\_ISL\_733592 and EPI\_ISL\_733681, the heatmap had lighter colors compared to the rest of the figure. And there appear to be four distinct light-colored regions within the previously mentioned sample IDs. On the outer regions of the heatmap, there appear to be two distinct darker colored lines between samples labeled as EPI\_ISL\_733681 and EPI\_ISL\_741266. Lastly, another feature that might confuse people is the white-colored line that runs diagonally across the heatmap, from the upper left corner to the lower right corner of the heatmap.

The light-colored regions represent the high percentage of similarity between the two samples that are compared and the darker-colored regions represent the low percentage of similarity between the samples. Since the matrix of the heatmap is Mi,j, the white diagonal line only means the sample is identical to itself, which explains the 100% similarity. By observing the heatmap, there are roughly 1/10 of the samples that are outliers for the December 8th, 2020, SARS-CoV-2 genome, as their values had a low percentage of similarity to each other and other samples collected that day. While the heatmap from December 8th seems lighter in general compared to the other heatmaps, we can tell some of the mutated viruses might have already disappeared in the process of spreading. But if we look closely at how similar/different a sample can be compared to other samples collected in one day, we can find some other interesting features: comparing the samples from December 8 and the ones from November 3rd, we can see the darkest area in the November 3rd heatmap has a lower percentage of base similarity than the darkest spot from December 8th. Thus, the most different sample from November 3rd had more variants within than the one collected on December 8th. Lastly, all the heatmaps from November 3, 2020, to December 8, 2020, SARS-CoV-2 genome have grid-like patterns. There are large squares containing light-colored values near one corner of the heatmap and darker-colored values that formed stripes in the heatmap.

The heatmaps conclude that the spread of SARS-CoV-2 in England had fewer variants. From November 3 to December 8, the variants of the SARS-CoV-2 genome decreased because of more appearance in light-colored values in the heatmaps. The November genomes and December 5 genome had several dark-colored strips in the heatmaps which indicated a large variety of variants in the SARS-CoV-2 genome. In the latest genome file, December 8, the heat map had more light-colored values and fewer darker-colored values. This indicates fewer variants in the genome and the Alpha strain dominated the SARS-CoV-2 genome.

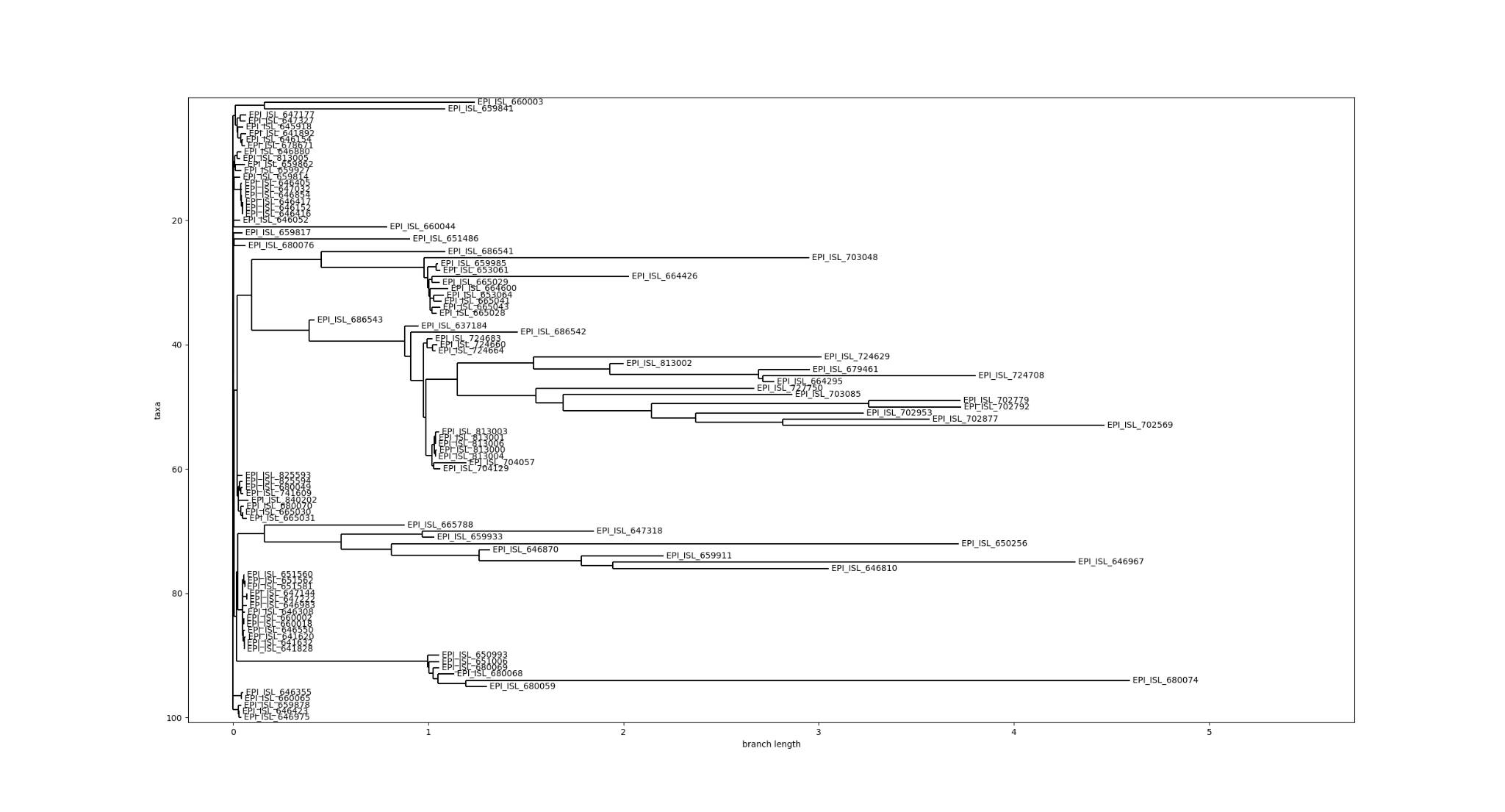


**Figure 2:** Heatmap of the December 8, 2020, SARS-CoV-2 genome

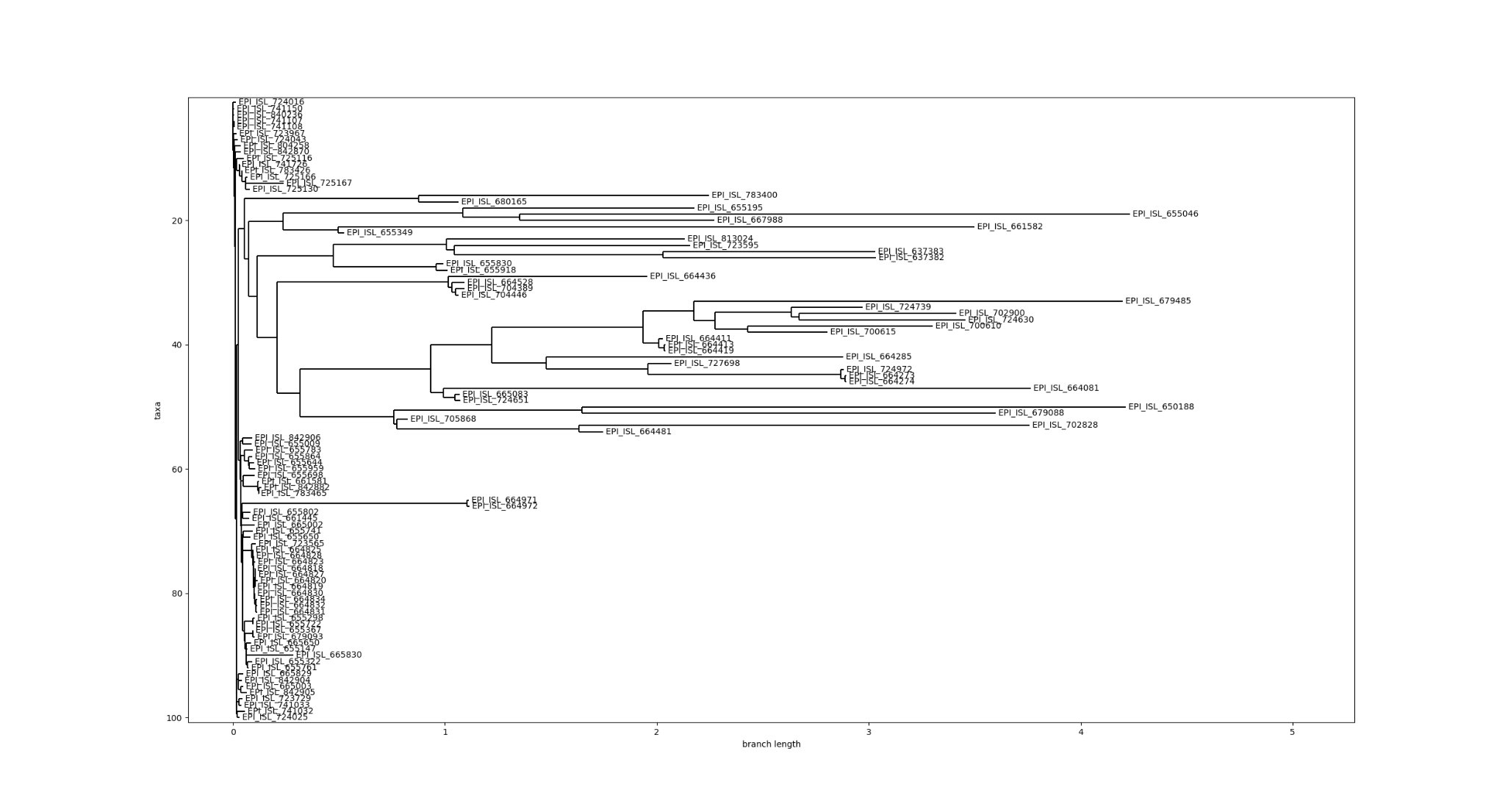
For part five of the lab, we answered question nine by writing a neighbor-joining function to create an evolutionary tree for the data collected on each day. In Figure 3, the tree shows evolutionary relations between the genome samples collected on November 3rd. Figure 4 shows the tree of the November 10th samples. Figure 5 shows the tree of the December 8th samples. The X-axis of the figure represents the length of each branch on the tree and the Y-axis represents the position of each taxa name stored in the input file. In Figure 3, the tree shows taxas connected by short branches which are closely related samples that don't have many variants, and long branches which indicate further relations and more mutations between its connected samples. As we can see from the trees, most branch lengths are lower than 1, which indicates the viruses are mostly similar to each other. But there are also long branches that can have a length value around 2-4, which indicates rapid mutations between the samples it connects. As we can see across different days, the number of outliers in the trees decreased but the branch length increased. This means that the number of varied samples decreased but the changes within each varied sample are greater as time passes.

To determine where the B.1.1.7 variants are on the tree, we used the labels produced from exercise 7, which are the IDs for the samples that satisfy all three features of the B.1.1.7 variant. By comparing the labels and the outliers on the tree from December 8th, we found that almost half (40%) of the outliers are considered as the B.1.1.7 variants we found in the former exercise. While other outliers on the tree might be other kinds of mutated samples, we can tell that a lot of the B.1.1.7 variants play a big part in this group of further relatives to other samples.

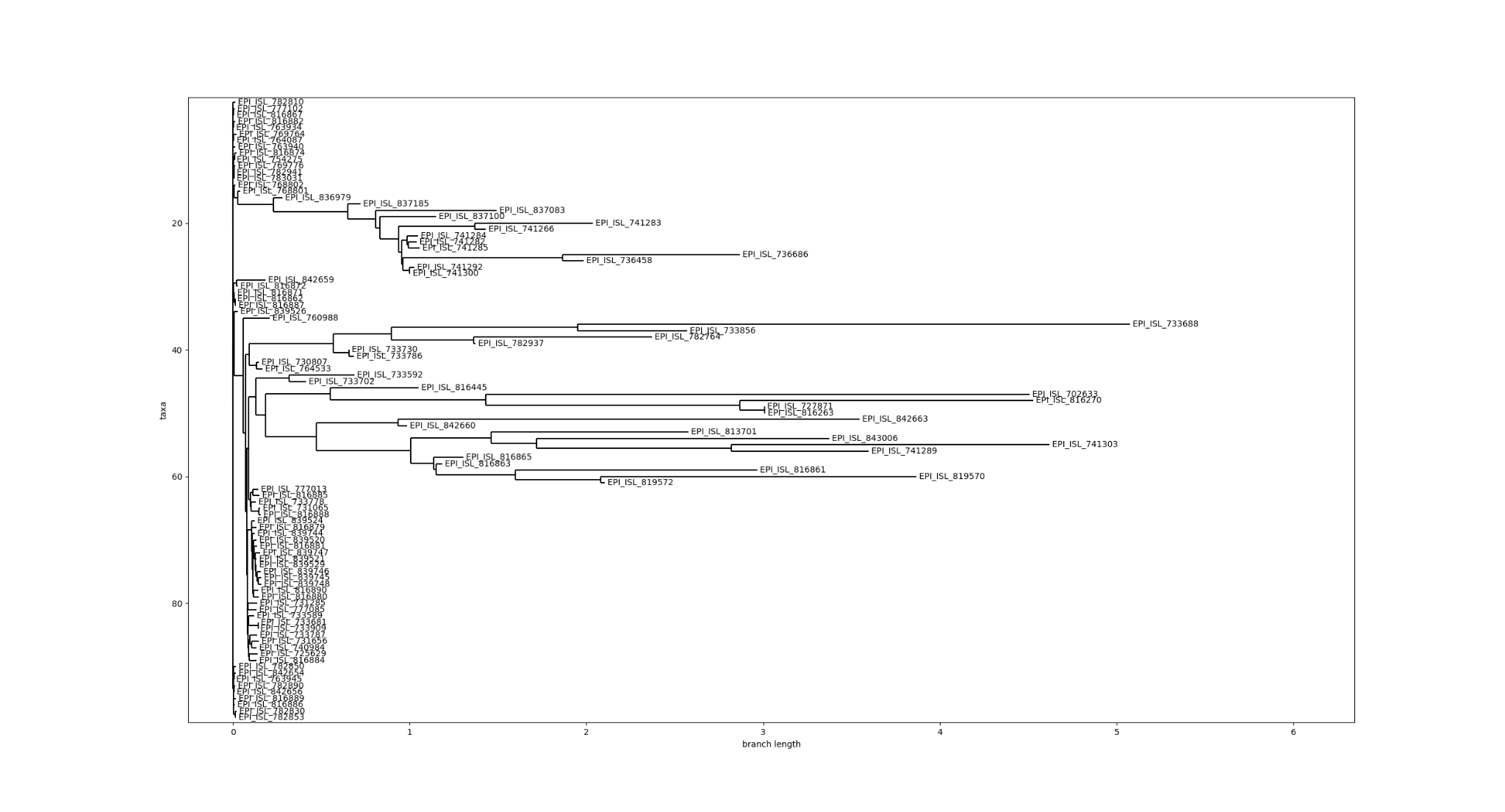
In conclusion, variants like the B.1.1.7 variants were found in the SARS-CoV-2 genome from November 3, 2020, to December 8, 2020. The variants of the SARS-CoV-2 genome in England changed over time. The SARS-CoV-2 genome had fewer variants as time passed because certain types of rigidly mutated variants dominated the SARS-CoV-2 genome.



**Figure 3:** November 3, 2020 SARS-CoV-2 genome



**Figure 4:** November 10, 2020 SARS-CoV-2 genome evolutionary tree



**Figure 5:** December 8, 2020 SARS-CoV-2 genome evolutionary tree

**Bibliography:**

Compeau, and Pevzner. Bioinformatics Algorithms: An Active Learning Approach, Active

Learning Publishers, 2015

NCBI. “NCBI Multiple Sequence Alignment Viewer 1.21.0.” *National Center for*

*Biotechnology Information*, U.S. National Library of Medicine,

https://www.ncbi.nlm.nih.gov/projects/msaviewer/.