Title: SARS-CoV-2 Spike Protein Phylogenetic Variability in US Geographical Regions and Time Periods.

Abstract/Summary paragraph

Coronavirus Disease 2019 is a newly emerging infectious disease currently spreading across the world. The disease is caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). A large number of spike proteins cover the surface of SARS-CoV-2 and bind to the host angiotensin-converting enzyme2 (ACE2). Once bound, virus entry to the host cell is promoted and the replication and transcription of the RNA composing SARS-CoV-2 can begin. Due to the crucial functions the spike protein plays in viral replication, it is a key target for COVID-19 vaccines and therapies. Here we show the relatively low genetic variation of the SARS-CoV-2 spike protein across time periods and various geographies in the United States, despite the sheer number of infected persons. Through analysis of 150 random samples collected from 15 US states, roughly 5 collected prior to July 2020 and 5 collected after, we found no significant outgroups of sequences when compared by collection time or geographic location. In addition, samples from China collected in early 2020 showed clustering with samples collected in the first period of analysis, suggesting that the genetic diversity found occurred in the US. Additional studies are needed to confirm such a suggestion.

Summary of previous findings

Previous research of the COVID-19 spike protein has revealed mutations that seem to be more dominant and could increase infectivity of the virus. One of these, the D614G mutation, was first discovered in a genome sampled in Europe and has quickly spread worldwide (Isabel et. al. 2020). This paper shows that there seems to be an increase in rates of samples having this mutation in locations where it is present, but does not have sufficient data to state this with confidence. This mutation enhances viral replication and animal models showed increased fitness in the upper airways which explains the apparent increased infectivity seen with this mutation (Plante et. al. 2020). An analysis of spike protein mutations in the Middle East and Africa showed there was a significant increase in the rates of samples containing the D614G mutation from 63.0% of samples in February 2020 to 98.5% in June 2020. These phylogenetic clusters were found using Maximum Likelihood and Bayesian methods to determine how the mutations increased with time (Sallam et. al. 2020). These phylogenetic analyses show that there seems to be a worldwide evolutionary convergence of the COVID-19 spike protein to have this mutation with higher replication.

Results

We randomly selected 150 sequences of the SARS-CoV-2 spike proteins from 15 US states, with an additional 7 samples collected in early 2020 in China as a reference outgroup.
For most states, 5 samples were selected preceding July 2020, and 5 from that date forward. Due to limit in sample availability, only pre-July samples were included for Indiana and Georgia, and only 1 out of 10 samples were selected in the post-July period for Florida and Illinois. Most pre-July samples (66%) were collected in the month of March. Post-July samples were more evenly distributed from July to October **(Figure 1)**. These sequences were aligned using MUSCLE.

Through the Akaike information criterion (AIC) approach, we determined that the mitochondrial models mtmam or mtArt would be the worst evolutionary models to calculate the distance matrix, while JTT and FLU (an influenza model) were the best models. These findings lead us to select the latter, FLU, for calculation of distance matrices and further phylogenetic analyses **(Figure 2)**.

We utilized three approaches to determine phylogeny clustering: UPGMA, NJ, FastME. Neither approach suggested strong clustering patterns by time period or US region. The UPGMA approach suggested a large number of sequences to be nearly identical, but with no differences in terms of time period or US

region **(Figure 3)**. However, the 7 outgroup samples from china were included in this cluster **(Figure 3A)**. This feature was not observed in the NJ or FastME approaches. In these two approaches, the outgroup samples were not clustered together.
Correlation analysis gave precedent to both NJ and fastME (both r=0.990561) over UPGMA (r=0.815507).

Conclusion

We analyzed 150 samples with the goal of assessing patterns in SARS-CoV-2 spike protein phylogeny by time (pre- vs post-July 2020) and US geographical region. Clustering by either variable was not observed with either phylogeny approach (UPGMA, NJ, FastME). UPGMA clustered the outgroup of samples (China, early 2020) with a large group of other samples. This could suggest that genetic variation occurred largely within the US, without geographical constraints, in the latter two thirds of 2020.
Further analysis with a larger sample size and a stronger selection methodology are warranted to confirm our suggestive findings and answer further research questions.

Methods

We aligned our sequences using EMBL-EBI's MUSCLE tool and the default parameters. Using the software R, we read in the aligned amino acid sequences and converted the alignment to a phyDat object. PhyDat objects are primarily used as an input for the tree building functions in the phangorn package, though other methods for tree building could also be used. To determine which evolutionary model was the best fit both the lowest AIC and lowest BIC were piped to the phangorn::dist.ml function to create a distance matrix. For brevity, only the FLU model was chosen, though there were several nuclear models that were also considered "best" (see: "AIC measures" figure).

For tree construction, three distance-based methods were compared: Neighbor-Joining (NJ), UPGMA, and fastME. The following were considerations for model inclusion: evolution rate; popularity; and efficiency. All three methods were implemented using phangorn's built-in functions and were run using the default settings. To visually compare methods, we created plots using both regional data (see "Model selection based on US regions" plot) and periodic data (see "Model selection based on time intervals"). To statistically compare methods, we ran correlation tests between the distance matrix and the distances within each methods' tree.

References

Sallam, M., Ababneh, N., Dababseh, D., Bakri, F., & Mahafzah, A. (2020). Temporal increase in D614G mutation of SARS-CoV-2 in the Middle East and North Africa: Phylogenetic and mutation analysis study. doi:10.1101/2020.08.24.20176792

Isabel, S., Graña-Miraglia, L., Gutierrez, J.M. et al. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. Sci Rep 10, 14031 (2020). https://doi.org/10.1038/s41598-020-70827-z

Plante, J.A., Liu, Y., Liu, J. et al. Spike mutation D614G alters SARS-CoV-2 fitness. Nature (2020). https://doi.org/10.1038/s41586-020-2895-3

Figures

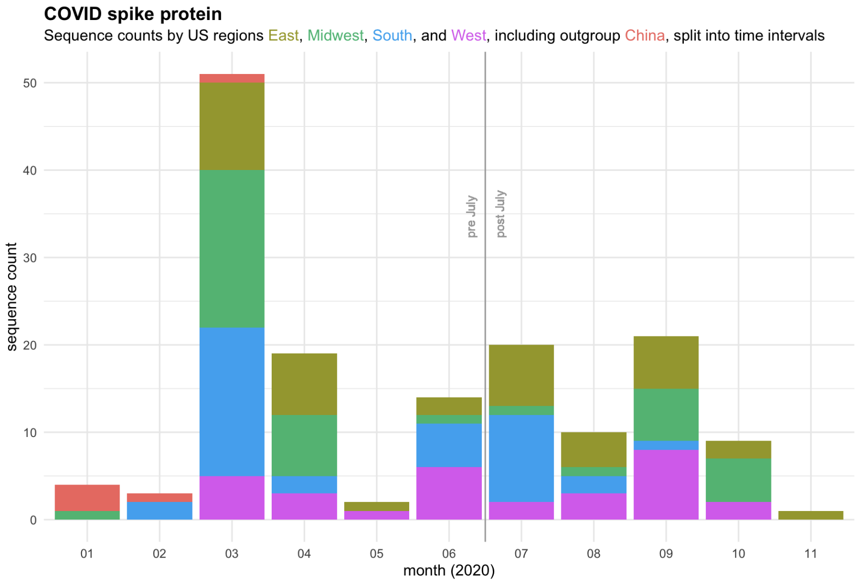

Figure 1: Distribution of SARS-CoV-2 spike protein samples (n=157) selected by month and geographical region (including outlier group).

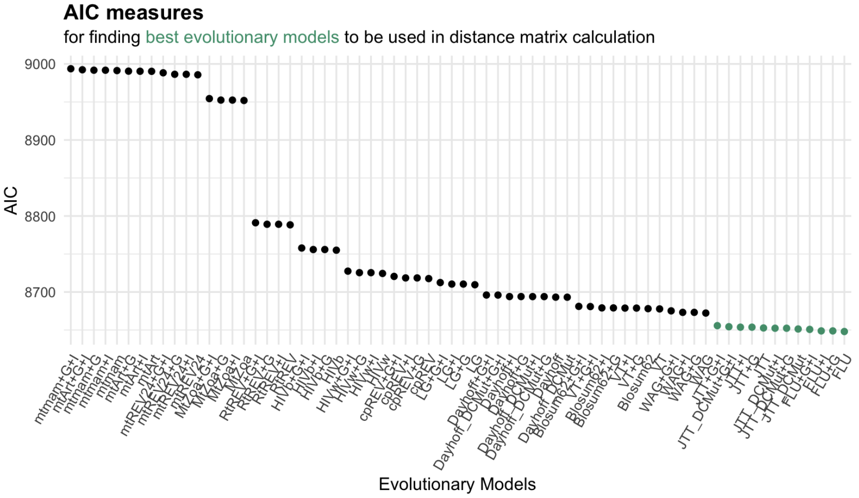

Figure 2: Akaike information criterion (AIC) for evolutionary models utilized in distance matrix calculation.

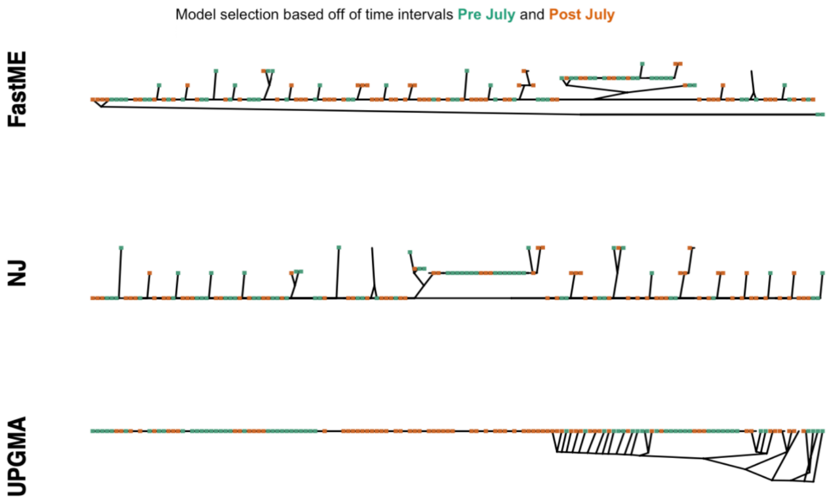
Model selection based off of time intervals Pre July and Post July
FastME
NJ
UPGMA

Figure 3A: Phylogenetic clustering of SARS-CoV-2 spike protein samples (n=157) utilizing FastMe, Neighbor Joining (NJ) and UPGMA, labeled by period of sample collection (pre- vs post-July 2020).

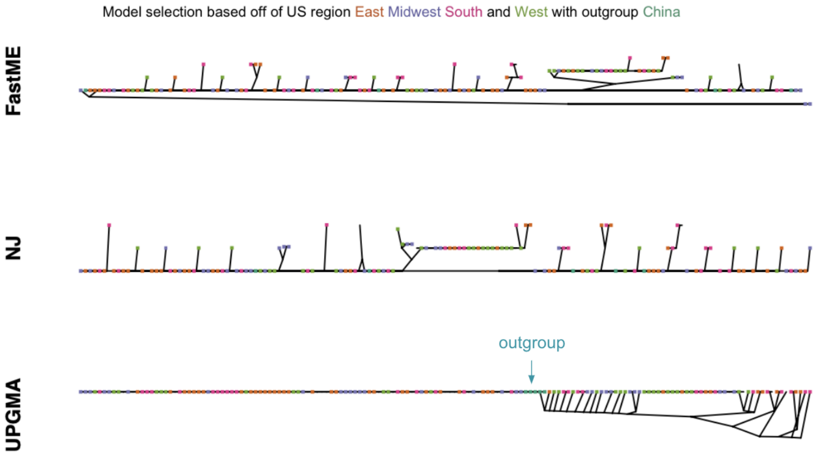
Model selection based off of US region East Midwest South and West with outgroup China
FastME
NJ
outgroup
UPGMA

Figure 3B: Phylogenetic clustering of SARS-CoV-2 spike protein samples (n=157) utilizing FastMe, Neighbor Joining (NJ) and UPGMA, labeled by geographic region (US East, Midwest, South, West, and outgroup - China).