Katherine Duseau
Class of 2022
Genetics and Biotechnology

Genome Analysis of SARS-CoV-2 Virus Variance: Identification and Assessment

Advisor: Alireza G. Senejani Ph. D.
Department of Biology and Environmental Science

During the last 2 years since late 2019, the world has been devastated by the Severe Acute Respiratory illness coronavirus (SARS-CoV-2), the driving force behind the current COVID-19 pandemic (WHO, 2021). Like all organisms on earth, viruses are continuously mutating and evolving becoming more virulent as time goes on. Identifying these mutations is important to combat viruses like SARS-CoV-2 and develop effective vaccines and treatments.

The purpose of this study was to identify novel mutations in coronavirus that have the potential to impact, transmission, vaccine efficacy, and virulence. To conduct this study, reference and sample sequences were collected from the NCBI GenBank database and analyzed through the SnapGene alignment program. This study focused on comparing sequences from a single lineage of concern, the delta lineage. It was determined that this lineage would be the focus of the study as it has been found in breakout cases in vaccinated populations (Meseret, B. 2021).

The SARA-Cov-2 virus is approximately 30 thousand bases (kb) long with 10 open reading frames coding for 4 functional proteins. The functional or structural proteins that are coded for include the M(membrane) E(envelope) N(nucleocapsid) and S(spike), each protein codes for a physical structure that make up the body of the virus (Naqvi, 2020). The Spike protein regulates through interactions with the ace2-receptor controlling entry to the host cell and transmission (Barnes, C. 2020). In addition, the spike protein houses a significant portion of the virus's epitope regions. Epitope regions are portions of the virus that can be recognized by the host's antibodies that target the virus for destruction or stop it from entering and infecting host cells. It is for this reason that the primary genome region analyzed in this study codes for the Spike Protein.

The sequences analyzed in this study were collected between June-July 2021. In total, 88 sequences were collected from 18 of the 50 states, including Massachusetts, New York, Connecticut, Florida, Georgia, Alabama, Ohio, Colorado, Utah, Maryland, Arkansas, Kansas, Nevada, Arizona, Texas, California, & Washington. Notably, the sequence reports collected from Massachusetts on July 7th & 8th had large numbers of unidentifiable nucleotides within the spike gene and were not considered when comparing significance. From these sequences, 15 novel mutations were identified when compared to the reference sequence collected from India in April 2021. The significance of these mutations was hypothesized by the position found in the genome. Ten of the 15 identified mutations fall within epitope regions along the spike gene. All ten significant mutations identified were caused by single nucleotide changes that result in a missense mutation; those that lead to amino acid changes and were identified in multiple samples. The presence of the same mutation in multiple samples suggests that the identified changes were not due to sequencing error. Additionally, four mutations V70F, K77T, T95I and, S112L were found to fall within regions of overlapping epitopes. Interestingly, each of these mutations were found in close succession in the N-terminal domain of the spike gene. This is significant because although the receptor binding domain (RBD) physically contacts the ace2-receptor regulating cell entry, the N-terminal domain was found to be a hotspot for single and overlapping epitopes or sequences that can be recognized by both B-cells and T-cells (Yi, C. 2020; Fatoba, A. 2021; Li, y. 2021).

The positioning of these mutations suggests that each could have a significant impact on the way antibodies recognize the virus, potentially impacting transmission, vaccine efficacy, and/or virulence. Further research is needed to investigate the relationship between the positioning of the mutations and the way antibodies recognize the virus, including the implications of these results for understanding transmission, vaccine efficacy, and virulence.

References

Barnes, C. et al., (2020). Structures of human antibodies bound to SARS-CoV-2 spike reveal common epitopes and recurrent features of antibodies. *bioRziv,* preprint, DOI: 10.1101/2020.05.28.121533

Fatoba, A. et al., (2021). Immunoinformatic prediction of overlapping CD8+ T-cell, IFN-c and LI-4 inducer CD4+ T-cell and linear B-cell epitopes based vaccines against COVID-19 (SARS-CoV-2). *Vaccine,* 39(7), 1111-1121, https://doi.org/10.1016/j.vaccine.2021.01.003

Li, Y. et al., (2021). Linear epitope landscape of the SARS-CoV-2 Spike protein constructed from 1,051 COVID-19 patients. *Cell Reports,* 34(13), DOI:10.1016/j.celrep.2021.108915

Meseret, B. et al., (2021). COVID-19 Vaccine Breakthrough Infections Reported to CDC-United States, January 1-April 30,2021. *MMWR,* 70(21), DOI:10.15585/MMWR.MM7021E3

Naqvi, A. et al., (2020). Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomic approach. *Molecular Basis of Disease,* 1866(10), DOI:10.1016/j.bbadis.2020.165878

World Health Organization. (2021). *WHO coronavirus (COVID-19) Dashboard*. World Health Organization. https://covid19.who.int/.

Yi, C. et al., (2020). Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. *Cellular and Molecular Immunology,* 17(6), 621-630, https://doi.org/10.1038/s41423-020-0458-z