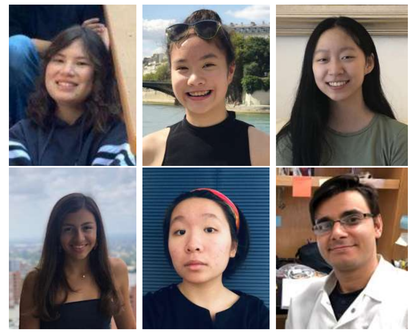
OMICS RESEARCH SYMPOSIUM

Bioinformatics Pipeline to Calculate the Frequency of SARS-CoV-2 Variants of Concern Versus Time and Web App Visualization

AneBritt Borchert, Angelina Choy, Angela Jiao, Lucia Losada Terreros, Kira Tang, Dr. Gepoliano Chaves



Introduction

The spike proteins on SARS-CoV-2, the virus that causes COVID-19, interact with ACE2 receptors on the cell, prompting the cell to envelop the virus in an endosome and bringing it into the cytoplasm.

Once inside, SARS-COV-2 releases its positive sense RNA for the cell's ribosomes to translate into additional functional viral proteins. The rest of the DNA is also replicated by DNA polymerase. The new viral proteins and genome then come together to form mature virions, which are then exocytosed back into the body, leading to a viral infection.

As COVID-19, caused by SARS-CoV-2, continues to wreak havoc across the globe, emerging viral mutations pose a great threat to the general public.

Once viruses replicate, mistakes may be made in the gene replication process, causing a Single Nucleotide Polymorphism (SNP) [Shen 1999]. A SNP changes the resulting primary structure (amino acid sequence) of the protein, which can alter its structure and thus, its function. A buildup of these SNPs and may lead to spike proteins that could more easily bind to ACE2 receptors. A variant of concern (VOC) arises when the spike proteins of a viral strain exhibit this significantly higher binding affinity to ACE2 receptors.

SNP alterations in the genetic code of SARS-CoV-2 can be investigated using computational methods or pipelines, which use text files derived from DNA sequencing.

DNA sequencing is beneficial because it allows detection of mutations within the SARS-CoV-2 genome. At the same time, studying genetic alterations in the genome of SARS-CoV-2 provides a platform for understanding genomics and computational biology concepts in more complex genomes such as the human genome. In humans, bioinformatics pipelines can be used to study diseases such as such as cancer, using pipelines similar to pipelines used for SARS-CoV-2. Finally, computational pipeline used in this project can be used in the development of a web application for visualization of change in frequency of variants of concern of SARS-CoV-2, making our pipeline and web application a useful tool in bioinformatics for social innovation.

Through our work, we seek to understand the viral evolution of SARS-CoV-2 and its VOCs over various geographic regions using SNP identification in a three-step in-silico pipeline. We look to analyze the functions of each mutation and their effects on the virus' efficacy and transmissibility. To improve accessibility to this and related information, we also develop a web application for the visualization of a user-inputted set of sequences.

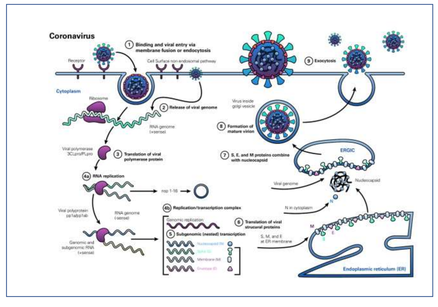


Figure 1: How COVID is able to reproduce and create more virions

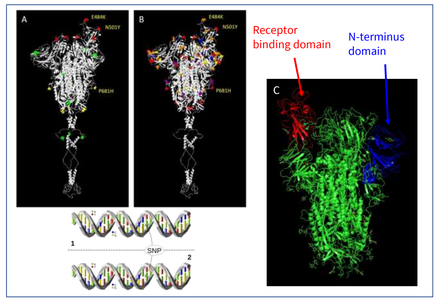


Figure 2: Spike protein model showing locations of the mutation. Image C generated using Pymol.

Methods

Overview
We looked at a total of 10 different countries, along with the strains from the Chinese rufous horseshoe bat (*R. sinicus*) and the Sunda pangolin (*M. javanica*). Namely, the 10 countries were Brazil, China, Germany, Japan, Mexico, Malaysia, Russia, South Africa, Spain, and Sweden. These were then analyzed across three different temporal ranges, i.e. July - December 2020, January - February 2021 and March - April 2021. We chose our data from a broad range of locations and collection dates across the globe in order to get a varied sample. Furthermore, by looking at the data of different time ranges, we were able to see the progression of the virus and its virulence.

The Pipeline
The data was gathered through the Global Initiative on Sharing Avian Influenza Data (GISAID). We used RStudio to run the code and generate the pipeline. Specifically, we ran all the sequences through SAMtools and BCFtools, which are used for variant calling and converting the original data files into VCF files.

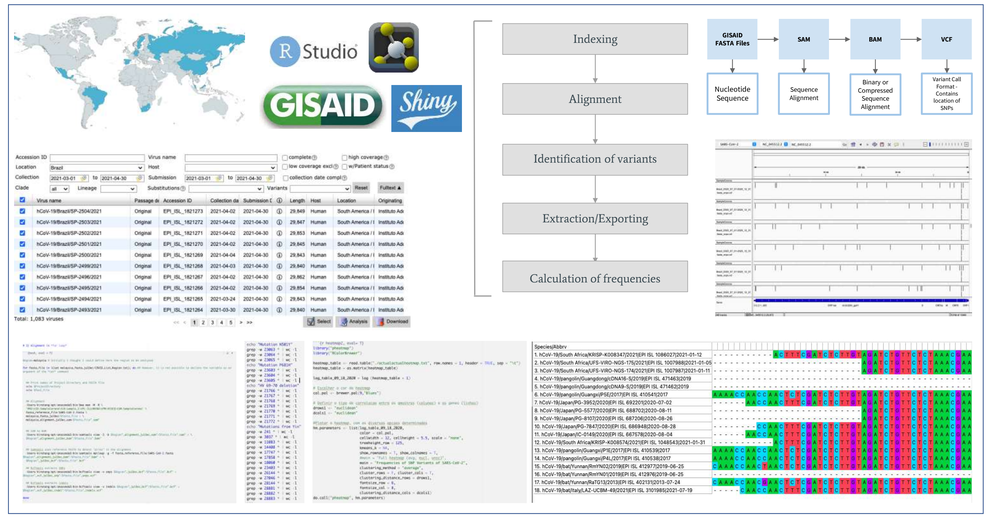
The general organization of the pipeline has five main steps. In order, they are:

1. Indexing, done through BWA (same index for different samples, analogous to an index in a book)
2. Alignment
3. Identification of variants through comparison with the reference file (a sequence of virus collected from one of the first patients diagnosed with COVID-19)
4. Extraction/Exporting
5. Calculation of Frequencies (for 27 specific mutations)

Hierarchical Clustering
Creating a heatmap allowed us to better visualize the frequencies of the SNPs through varying levels of color intensity. In order to construct the plot, the master spreadsheet with all countries – including bat and pangolin – is exported as a .txt file and put into the code. A logarithmic scale was used to homogenize the frequency of the variants and normalize them.

VCF Files Visualization
We used a genome browser called IGV to visualize Variant Call Format (VCF) files. This allowed us to see how mutations are distributed across an entire genome. Furthemore, we were able to make comparisons between mutation distribution in different time periods and geographical regions.

Phylogenetic Tree
We used MEGAX (Alignment by ClustalW) to align the viral sequences and create a phylogenetic tree diagram that represents the relatedness between the sequences in our data set.



Web Application Development
We used the Shiny library in R to construct our web application for simplicity. The app accepts user input of a DNA or protein sequence file in the .fasta format, which it uses to construct a frequency table and phylogenetic tree. The frequency table is the result of multiple sequence alignment and displays the number of sequences in each position that have a certain nucleotide or amino acid, and the phylogenetic tree is formed through the neighbor joining algorithm. Limitations in this method arise from the fact that it only produces one tree, but it was selected for its speed, which is valuable in an online application, and generally better performance than UPGMA [Saitou 1987]. The app was tested with SARS-CoV-2 sequence data from GISAID.

Results

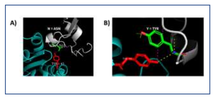


Figure 4: Pymol N501Y mutation visualization

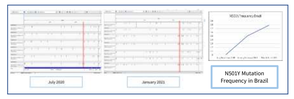


Figure 5: Genome browser visualization of VCF files. Highlighted frequency of mutation N501Y in Brazil.

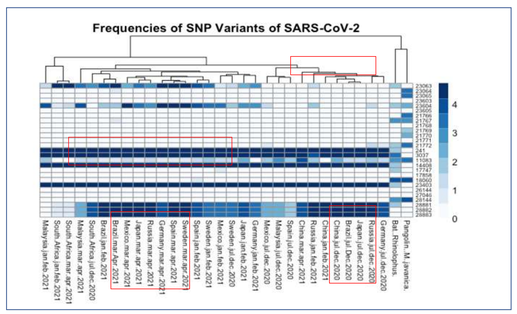


Figure 6: Heatmap of each of the frequencies per country

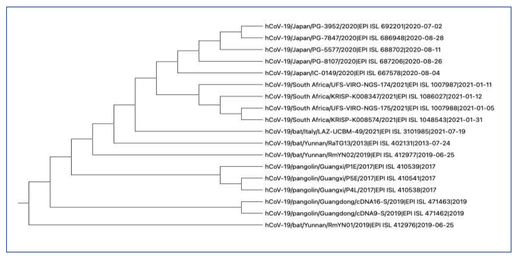


Figure 7: Phylogenetic Analysis of Pangolin, Bat, Japan and South Africa

Pymol Analysis of N501Y Mutation
We used the Pymol software to visualize the SARS-CoV-2 Spike Protein. We were able to analyze samples with and without the N501Y mutation, which is present in variants from the UK, South Africa, and Brazil. From these images, we concluded that the mutation leads to an increased number of polar interactions (hydrogen bonds), which increases the affinity between the spike protein and the ACE2 receptor. This indicates that transmiitibity could potentially be increased.

VCF File Visualization in Genome Browser
We created a chronological comparison using mutation N501Y (base pair 23063 in spike protein). Figure 5 shows the data from Brazil in both July 2020 and January 2021. The section highlighted in red shows that none of the sequences sampled in July 2020 had mutation N501Y. However, 3/7 of the mutations from January 2021 do include this mutation. Based on this information and other samples we created a frequency chart of the mutation in Brazil from July 2020 to April 2021. From this chart we can predict that an even greater percentage of sequencing done today includes this mutation, and the trend will continue to rise.

Phylogenetic Analysis and the Heatmap
When looking at the overall trend of the SNP frequencies, it's clear that as time progresses, certain SNPs became more prominent. The most notable being 23063 and 23604 which belong to mutations N501Y and mutation P681H, respectively. The fold change between them was significantly high across the majority of countries as it neared towards March 2021. SNPs 241, 3037, 14408, and 23403 stayed relatively the same whereas 28881, 28882, 28883 seemed to increase in tandem with one another. This likely had an effect on fidelity.
Looking at a subset of the data (Figure 6), a number of the SNP frequencies, while apparent in the human genome, are absent in that of the pangolin and bat. It highly suggests that one of these particular SNP frequencies could be responsible for the zoonotic jump between the species. Based on each of the SNPs, it's likely that this was a result of 23403. As it's located on the spike glycoprotein, 23403 is an important component in determining host range.
As is seen in the heatmap, as the virulent spread over time, the genetic variation increased, leading to greater separation between countries. However, Bat and Pangolin had formed their own cluster and differed the most from the other countries, only further showing the variation between each virulent strain. This shows the rate at which the virus is able to mutate and develop characteristics to be able to increase its transmissibility and efficacy.
Within the phylogenetic tree, the bats themselves have their own cluster- further supporting the heatmap, with one outlier at the bottom. This outlier could possibly be a representation of the epicenter or a separate strain as well. SNP variants are more closely related to the virus found in bats than the virus found in pangolins.

Web Application
We produced a working web application (Figure 9). The frequency table displays the frequencies for the first 10 positions in the DNA sequence and tracks the number of sequences that have an A, C, G, T, or deletion at each position. Similar results were obtained for the amino acid test data, and a phylogenetic tree was visualized. This allows for quick analysis of recent data and simplifies the comparison of sequences for the viewer.

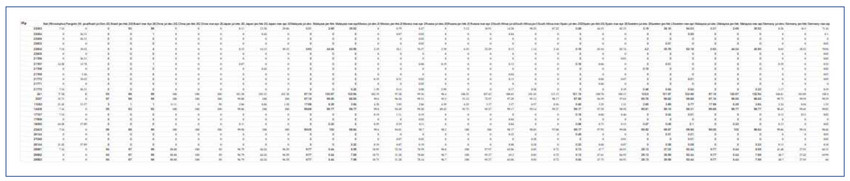


Figure 8: Master spreadsheet with all SNP frequencies across each respective country

Conclusions

Across various countries, the general trend is that over time, greater genetic variation occurs within the SNPs. Most countries with data between March-April 2021 included a large amount of mutations in comparison with the genome sequences from July-December 2020. There overall was more clustering between each of these countries as a result. The SNPs which remained the same (i.e. 241, 3027, 14408, and 23403) seemed to be necessary to the transmissibility of the virus itself, particularly when combined together. One commonality that we found was that the SNPs which had higher frequencies tended to affect the spike proteins on the surface of the virus. In doing so, it helps the virus to be able to bind more effectively to the ACE-2 receptors. Thereby, further increasing its ability to overtake host cells.

The developed web app can be an easily accessible tool for health policymakers to visualize our findings, which can guide future actions against the pandemic.

Future Research
In order to continue and build on the study, we would want to continue tracking the evolution of SARS-CoV-2 variants as data from around the world continues to be collected. We further want to explore various mutations and their impact on viral abilities such as zoonotic jumps between species. This is especially vital as mutations often are responsible for more virulent variants. Being able to track its evolution helps us understand the threat to the general public.

For the web app, we plan to complete a line graph component for the visualization of the change in variant frequencies over time using dated sequence files from our pipeline.

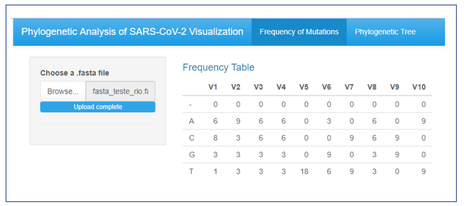


Figure 9: Web application UI

Published Bibliography

Collins, A., C. Lonjou, and N. E. Morton. "Genetic Epidemiology of Single-Nucleotide Polymorphisms." Proceedings of the National Academy of Sciences of the United States of America 96, no. 26 (1999): 15173-5177. Accessed August 14, 2021. http://www.jstor.org/stable/121226.

Gostin, Lawrence O., Alexandra Phelan, Michael A. Stoto, John D. Kraemer, and K. Srinath Reddy. "Virus Sharing, Genetic Sequencing, and Global Health Security." Science 345, no. 6202 (2014): 1295-296. Accessed August 14, 2021. https://www.jstor.org/stable/24917590.

Loverdo, C., Park, M., Schreiber, S., & Lloyd-Smith, J. (2012). INFLUENCE OF VIRAL REPLICATION MECHANISMS ON WITHIN-HOST EVOLUTIONARY DYNAMICS. Evolution, 66(11), 3462-3471. Retrieved August 19, 2021, from http://www.jstor.org/stable/23273884

Padan, Carmit. Report. Institute for National Security Studies, 2020. Accessed August 14, 2021. http://www.jstor.org/stable/resrep25527.

Pulliam, Juliet R. C., and Jonathan Dushoff. "Ability to Replicate in the Cytoplasm Predicts Zoonotic Transmission of Livestock Viruses." The Journal of Infectious Diseases 199, no. 4 (2009): 565-68. Accessed August 14, 2021. http://www.jstor.org/stable/40254458.

Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987 Jul;4(4):406-25. doi: 10.1093/oxfordjournals.molbev.a040454. PMID: 3447015.

Shen, Ling X., James P. Basilion, and Vincent P. Stanton. "Single-Nucleotide Polymorphisms Can Cause Different Structural Folds of MRNA." Proceedings of the National Academy of Sciences of the United States of America 96, no. 14 (1999): 7871-876. Accessed August 14, 2021. http://www.jstor.org/stable/48393.

Shendure, J., Balasubramanian, S., Church, G. *et al.* DNA sequencing at 40: past, present and future. *Nature* 550, 345–353 (2017). https://doi.org/10.1038/nature24286

Virus Cell Reproduction. (1954). The Science News-Letter, 65(3), 42-43. https://doi.org/10.2307/3933464

V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., & Thiel, V. (2020). Coronavirus biology and replication: Implications for SARS-CoV-2. Nature Reviews Microbiology, 19(3), 155-170. https://doi.org/10.1038/s41579-020-00468-6

Wong, S., & Yuen, K. (2005). Commentary: Zoonotic Potential Of Emerging Animal Diseases. BMJ: British Medical Journal, 331(7527), 1260-1260. Retrieved August 19, 2021, from http://www.jstor.org/stable/25455500

World Health Organization. (2021). (Rep.). World Health Organization. Retrieved August 19, 2021, from http://www.jstor.org/stable/resrep33250

World Health Organization. Report. World Health Organization, 2021. Accessed August 14, 2021. http://www.jstor.org/stable/resrep30163.

Yin, C. (2020). Genotyping coronavirus SARS-CoV-2: Methods and implications. Genomics, 112(5), 3588-3596. https://doi.org/10.1016/j.ygeno.2020.04.016

Mentors: Dr. Gepoliano Chaves, Prof. Danilo Coelho, Dr. Jaime Amorim, Dr. Esther Camilo