SARS-CoV-2 Variant Analysis

Henrik Torres[1], Amith Vasantha[2], Helen Chow[3], Gepoliano Chaves PhD[4]

[1]Choate Rosemary Hall, 333 Christrian Street, Wallingford, CT 06492
[2]Basis Independent Silicon Valley, 1290 Parkmoor Ave, San Jose, CA 95126
[3]Lowell High School, 1101 Eucalyptus Dr, San Francisco, CA 94132
[4]University of Chicago, 5801 S Ellis Ave, Chicago, IL 60637

***Abstract*— The SARS-CoV-2 virus started the novel coronavirus pandemic. SARS-CoV-2 is an RNA virus that causes infection through the binding of the virion's spike protein to a cell's ACE2 receptor. The SARS-CoV-2 virion cleaves its way into the cell and deposits its RNA genome that hijacks the cell's RNA replication system to produce more virions. During replication, genetic variance arises through single nucleotide polymorphisms (SNPs) that can enable a zoonotic jump or affect the transmissibility or lethality of a virus [3]. Our research focused on studying these SNPs from collected FASTQ and FASTA files of human, pangolin, and bat SARS-CoV-2 genomes on online databases such as the NCBI SRA Browser and GISAID and running files through variant call pipelines. Our results confirmed SNP frequencies at locations in the genome that matched those of Yin [1]. Genomic comparison of SARS-CoV-2 between the humans, bats, and pangolins showed a distinct mutation at location 23403 on the spike protein that was only present in humans, possibly a mutation that facilitated a zoonotic jump. Regional frequency analysis of collected samples showed regional clustering and similarities. The results from our research further existing knowledge of the SARS-CoV-2 virus and can be further expanded upon to create regional vaccines specifically tailored to mutations that affect certain protein mechanisms.**

I. INTRODUCTION

The coronavirus pandemic has caused an unprecedented public health crisis with the potential to worsen if mutations were to arise that increased transmissibility or infectivity. This novel coronavirus is named SARS-CoV-2, and it comes from a family of coronaviruses that cause a respiratory infection. Past coronavirus outbreaks have been SARS in 2003 and MERS in 2012 [1]. However, the spread of SARS-CoV-2 has caused a pandemic of proportions last seen only in 1918 with the influenza. Proliferated spread of SARS-CoV-2 has led to more than 22 million confirmed cases and 777,000 deaths as of August 18, 2020 [2]. SARS-CoV-2 is an RNA virus with a single-stranded positive RNA genome. Structural proteins of SARS-CoV-2 virions include the spike protein, envelope protein, membrane protein, and nucleoprotein. To infect a cell, it uses its spike protein, located on the outer membrane protein, to bind to a cell's ACE2 receptor. SARS-CoV-2 cleaves its way into a cell with this binding mechanism and deposits its genetic material to hijack the cell's RNA replication mechanisms [3].

Within replication mechanisms needed to produce more SARS-CoV-2 virions occur mutations known as single nucleotide polymorphisms (SNPs). Our research focused on identifying these frequent SNPs, understanding the regional differences of SARS-CoV-2 variants, and searching for possible variants indicating a zoonotic jump (the transmission of a virus between species).

II. METHODS

Data was collected from the NCBI SRA Database and the GISAID Database in the form of FASTQ files and FASTA files, respectively. FASTQ files and FASTA files both function to store genomic information, with the important distinction being that FASTA files store a nucleotide sequence, while FASTQ files add quality scores to each nucleotide. Several FASTQ and FASTA files were collected from NCBI and GISAID with each continent except Antarctica being equally represented. To analyze the FASTQ and FASTA files, we used a variant calling pipeline (a shell script) on the University of California, Santa Cruz Hummingbird Server. Our pipeline aligned each genome file with the established SARS-CoV-2 reference genome from the UCSC Genome Browser, NC_045512, and stored the results in a Binary Alignment Map (BAM) file using Picard; then, we used GATK (a software supplied by Broad Institute along with Picard) to find SNPs and store them in a Variant Calling Format (VCF) file using a tool named Picard; then Picard marked duplicates and constructed a BAM Index. From the BAM Index, GATK (a genomic analysis tool) realigns the indels, calls and filters variants, and produces a Base Quality Score Recalibration. This pipeline was used for analyzing Sequence Read Archive (SRA) files from the NCBI SRA Database.

For the FASTA files from the GISAID Database, a simpler version of the aforementioned pipeline was used that realigned reads through Burrows-Wheeler Aligner (BWA), converted BAM alignment of the files to Sequence Alignment Map (SAM) alignment through Samtools, created .vcf (variant call format) files with Samtools, and marked SNPs and indels with BCFTools.

III. DATA ANALYSIS

After running the pipelines on the UCSC Hummingbird Server, we gathered the results in a spreadsheet and calculated the frequencies and regional frequencies (Figure 1). We constructed a Meta-Analysis Table (data not shown) containing information on the region of origin of the virus, size of the sequencing file analyzed, when the database was the Gene expression Omnibus and sequencing strategy adopted. The purpose of this table is to keep stable information about the virus for future reference. We also constructed a Frequency Table which we used to compute frequencies of different viral genotypes per different geographic region. This table also allowed execution of the pipeline for extraction of the different genotypes present in the files analyzed.

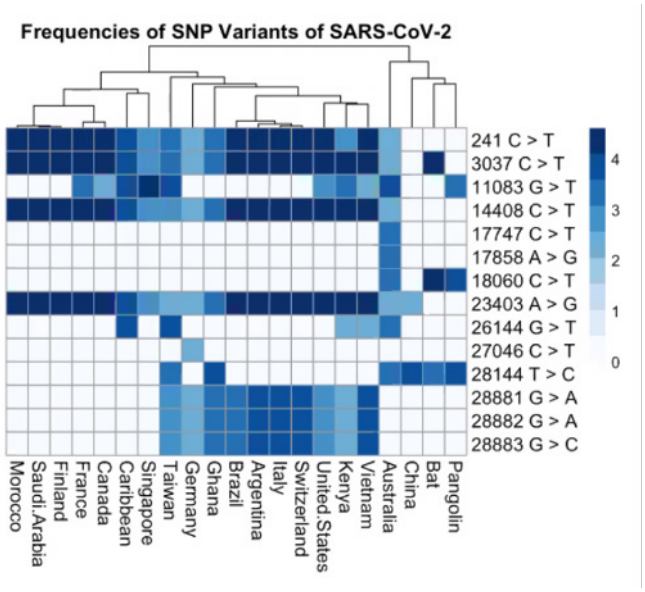

Figure 1. **Comparison of FASTA sequences from human samples in different regions around the world, bat samples, and pangolin samples.** The FASTA sequences were gathered from the GISAID Coronavirus Database and run through the variant call pipeline. The figure shows regional clustering based on the frequencies of SNP mutations. Samples taken from different countries become more genetically dissimilar as the virus traveled further from the epicenter.

IV. RESULTS AND DISCUSSION

The present COVID-19 pandemic highlighted the need for investing in understanding viral diseases, their spread, and the need for common hygiene practices across countries. Other measures that might become more important from here on is characterization of the genomic variants associated with severity of COVID-19 and incidence in certain geographic locations. Here, we compared FASTA sequences isolated from SARS-CoV-2 from across the globe. We found several overlaps between our own calculated frequencies and the frequencies found by Yin (2020). With these similar results, we were able to confirm their findings and take the research one step further by comparing the SARS-CoV-2 genomes of the samples from intermediate hosts to samples from humans. In Figure 1, we mapped the regional clustering of the different variants of SARS-CoV-2. As expected, they clustered regionally. As the virus travelled further from the epicenter, genetic variation increased and began to cluster regionally (Figure 1).

Additionally, we found that a mutation in position 23403 was present only in SARS-CoV-2 sequences infecting humans and not in bats or pangolins (Figure 1). This finding agrees with the notion that a different strain of the virus underwent mutation in order to infect humans. Furthermore, mutation in position 14408, which seems to be related to mutation 23403, was also identified in our study, present in all human sequences found infecting humans, except the Chinese samples (Figure 1). Additionally, we found that mutations in SARS-CoV-2's helicase (from SNPs in positions 17747 and 17858 of the genome) only occur in samples collected from Australia, and mutations in SARS-CoV-2's membrane glycoprotein (from SNPs in position 27046) only occur in samples from Germany (Figure 1).

Much is still unknown about SARS-CoV-2 and the disease it causes, COVID-19, but the research grows each day. Moving forward, further research can be conducted on how exactly SNPs in different regions of the genome alter protein structure and function. We have already found that the mutation at position 23404 may have facilitated a zoonotic jump, but more research must be done on this topic to confirm the claim. Additionally, more research must be done to determine why our heat map showed a closer clustering between bat sequences and human sequences because this may be an indication that they were not the original hosts of the virus. It is only a matter of time before all the SNPs and new arising SNPs are understood concretely. Given that there are more mutations in the genome of SARS-CoV-2 that fundamentally change how the virus operates in different regions, in the future scientists would be able to administer the most effective vaccines or antibody therapies for each region to prevent the spread of COVID-19.

V. ACKNOWLEDGEMENTS

We would like to acknowledge our mentor Dr. Gepoliano Chaves, PhD, for his endless guidance on this project and commitment to work. We would also like to thank the Science Internship Program (SIP) at the University of California, Santa Cruz for facilitating this research opportunity and providing a dedicated team for all participating interns. Additionally, we would like to acknowledge the NCBI SRA Database, the GISAID Database, and the submitting labs and hospitals around the world; this research would not have been possible without their work.

VI. REFERENCES

[1] 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

[2] WHO Coronavirus Disease (COVID-19) Dashboard. (n.d.). Retrieved August 31, 2020, from https://covid19.who.int/

[3] Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell, 181(2), 281-292.e6. https://doi.org/10.1016/j.cell.2020.02.058