College of ENGINEERING, COMPUTING AND APPLIED SCIENCES

Analysis of SARS-CoV-2 Genome Diversity

Jordyn Brooks[1], Abigail Thornton[2], Gaura Sinha[2], Congyue Annie Peng[3,4], Delphine Dean[3,4]

[1] Department of Genetics and Biochemistry, Clemson University [2] Department of Computer Science, Clemson University
[3] Department of Bioengineering, Clemson University [4] Research and Education in Disease Diagnostics and Intervention (REDDI)

REDDI LAB
Research and Education in Disease Diagnosis and Intervention

Abstract

The COVID-19 pandemic caused by SARS-CoV-2 has rapidly spread worldwide since its emergence in 2019. To better understand the genetic diversity of the virus, we aim to analyze the nucleotide diversity of SARS-CoV-2 in South Carolina using bioinformatics tools and techniques. By modifying pre existing pipelines and workflows in R and Python, we will generate a visual representation of the virus's genome that highlights potentially unstable regions where mutations occur more frequently. Using GISAID data, we will calculate nucleotide diversity and perform sequencing analysis to identify critical mutations associated with increased virulence and transmission. This analysis could provide valuable insights into the virus's evolution and spread, potentially contributing to the development of effective treatments and vaccines, as well as public health policy and response efforts.

Introduction

Genomic diversity and nucleotide diversity are crucial concepts in understanding the evolution and transmission of SARS-CoV-2. Genomic diversity refers to the genetic variation present in the entire genome of an organism, while nucleotide diversity estimates the variation in the DNA sequence at the nucleotide level. Both types of diversity can provide insights into the history, adaptation, and genetic structure of populations, and are particularly important in tracking the emergence and spread of new variants of the virus.

The genomic diversity of SARS-CoV-2 has been increasing over time, particularly in regions with high transmission rates. This diversity can be attributed to the accumulation of mutations over time, as well as the emergence and spread of new variants of the virus. Studies have shown that the diversity of SARS-CoV-2 genomes increased rapidly between January 2021 to December 2022, a period that coincided with the emergence and spread of several variants of concern, including Alpha, Beta, Gamma, Delta, and Omicron, indicating a high rate of mutation and adaptation.

The Diversity-GISAID was chosen in order to visualize data provided by the REDDI Lab here at Clemson due to its user-friendly and efficient software package for calculating nucleotide diversity in viral genomes. The program extracts data from a GISAID database, preprocesses it through editing the format, then calculates the variant replacement, frequency, and nucleotide diversity through time.

Materials and Methods

Results

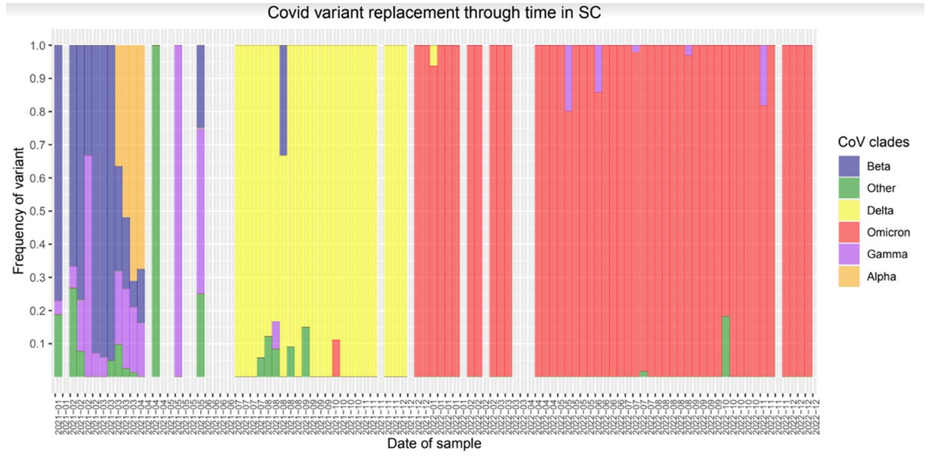


Figure 1: The frequency of each Covid variant over January 2021 to December 2022.

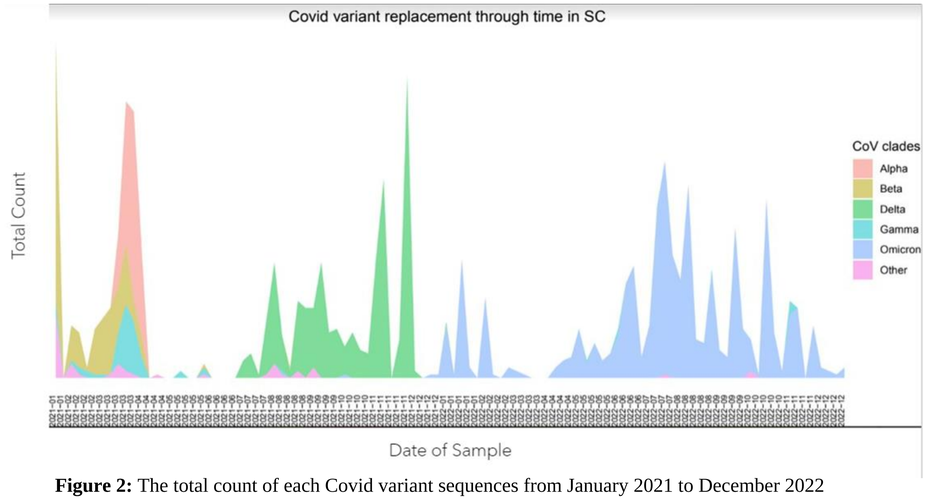
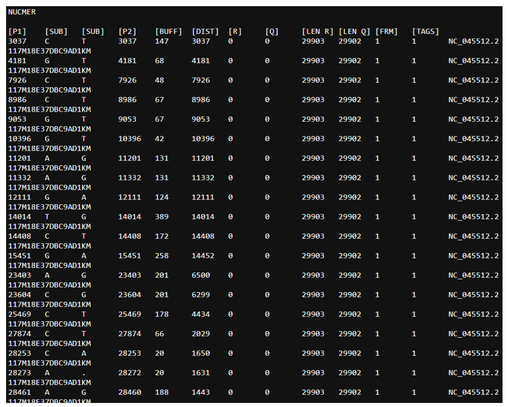


Figure 2: The total count of each Covid variant sequences from January 2021 to December 2022



NUCMER

[P1]	[SUB]	[SUB]	[P2]	[BUFF]	[DIST]	[R]	[Q]	[LEN R]	[LEN Q]	[FRM]	[TAGS]	
3037	C	T	3037	147	3037	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
4181	G	T	4181	68	4181	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
7926	C	T	7926	48	7926	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
8986	C	T	8986	67	8986	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
9053	G	T	9053	67	9053	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
10396	G	T	10396	42	10396	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
11201	A	G	11201	131	11201	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
11332	A	G	11332	131	11332	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
12111	G	A	12111	124	12111	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
14014	T	G	14014	389	14014	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
14408	C	T	14408	172	14408	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
15451	G	A	15451	258	14452	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
23403	A	G	23403	201	6500	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
23604	C	G	23604	201	6299	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
25469	C	T	25469	178	4434	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
27874	C	T	27874	66	2029	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
28253	C	A	28253	20	1650	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
28273	A	.	28272	20	1631	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												
28461	A	G	28460	188	1443	0	0	29903	29902	1	1	NC_045512.2
117M18E37DBC9AD1KM												

Figure 3: SNPs found between SARS-CoV-2 Wuhan reference genome and REDDI Lab test sample

Conclusions

This project uses the avaialbe data set on SARS-CoV-2 positivity generated through the CLIA certified clinical lab resides in REDDI, and the whole genome sequence data generated from randomly selected positive saliva samples, to explore the chronicle transition of the SARS-CoV-2 variant and its characteristic single nucleotide polymorphism (SNP) in the upstate South Carolina. We employed the MUMMER alignment tool, and the pipeline designed for the nucleotide diversity calculation along time by Dr. Schuyler Liphardt from the University of Montana. The preliminary analysis shows that using the current sampling strategy, we are able to determine the dynamics of the SARS-CoV-2 genome diversity through time. It is not surprising that the most diversity is found during the transition of the Alpha and Delta, when a time gap occurred between the dominant variants that gained advantages in transmission and infection. Although the individual count may be low from the strains that contributing to high diversity, the identified SNP, insertion, and deletion based on the global alignment of the virus genome provides evidence on the time frame of virus evolution. Additional analysis will be done to calculate the nucleotide diversity (π) and SNP frequency. Characterization of the genome diversity dynamics during the SARS-CoV-2 pandemic and its transition to endemic will provide crucial information and guidance on virus mitigation and prevention.

Future Work

- Use collected SARS-CoV-2 samples and the Wuhan reference genome to generate map of mutations and SNPs (Figure 3) via MUMmer and/or using R and Python.
- Continue working on calculating the nucleotide diversities of the covid variants through time through aligning sequences to SARS references.

Acknowledgements

We acknowledge Clemson University Creative Inquiry, and the assistance of Dr. Rooksana Noorai of the Clemson University Genomics and Bioinformatics Facility for services and facilities provided. The facility is supported by Grants P20GM109094 and P20GM139767 Institutional Development Awards (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health .

References

Arthur L. Delcher, Adam Phillippy, Jane Carlton, Steven L. Salzberg, Fast algorithms for large-scale genome alignment and comparison, *Nucleic Acids Research*, Volume 30, Issue 11, 1 June 2002, Pages 2478–2483, https://doi.org/10.1093/nar/30.11.2478

Liphardt. 2022. diversity-GISAID. https://github.com/liphardt/diversity-GISAID.git. (2023)

Mercatelli, D., & Giorgi, F. M. (2020). Geographic and genomic distribution of SARS-COV-2 mutations. https://doi.org/10.20944/preprints202004.0529.v1