Sequencing Sars-CoV-2 Variants of Concern to Find the Rate of Substitution

Authors: Jenny Harston, Jocel Clark, and Nathasya Asnawi

Affiliations:

University of Washington Bothell, School of STEM, Division of Biological Sciences

Keywords:

Covid-19, Sars-CoV-2, mutation, variant

Abstract

COVID-19 is a part of the coronavirus family and the emergence of the new Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) that has caused a global public health crisis. The novel coronavirus was first discovered in Wuhan China on December 31, 2019 and was labeled a pandemic in early 2020. With the virus quickly evolving, variants with increased infectivity have prompted the research of mutation rates in genomic sequence data of the different variants. Knowledge of the rate of mutation is important for tracking the spread of the virus, studying its phylogeny, and vaccine design and production. The Center of Disease Control (CDC) and World Health Organization (WHO) labeled eight variants of interest out of thousands of variants. To understand the virus better, we took one accession of these eight variants, parsed the FASTA files of their nucleotide sequences including the and the original virus of SARS-CoV-2, and aligned them using Multiple Sequence Comparison by Log-Expectation (MUSCLE) to obtain the rate of substitution and concluded that the rate of substitution differs between each virus variant.

Introduction

The Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) virus which is responsible for the Coronavirus Disease 2019 (COVID-19) has caused an ongoing global pandemic. The virus is evolving rapidly resulting in mutations and new variants. Spratt et al. (2021) indicates that new variants have greater infectivity compared to the original strain of the virus (Wuhan-Hu-1), which poses a great risk to overall global health. These new variants of SARS-CoV-2 virus across the world can be tracked due to genomic sequence data samples taken from infected individuals, which are available through resources such as the NCBI SARS-CoV-2 Data Hub and GISAID. These sequences have allowed scientists to study the genomic evolution of the SARS-CoV-2

virus by looking at their mutations. Understanding the process of mutations is important for tracking the spread of the virus and vaccine design (Maio et al., 2021, Spratt et al., 2021).

For our study, we collected eight different variants of SARS-CoV-2 nucleotide sequences, and one reference sequence, the Wuhan-Hu-1 strain that originated from China. The variants that we look at were B.1.1.7 (Alpha variant) from the UK, B.1.351 (Beta variant) from South Africa, P.1 (Gamma variant), from Brazil, B.1.617.2 (Delta variant) from India, B.1.617.1 (Kappa variant), B.1.526 (Iota variant), B.1.525 (Eta variant), and C.37 (Lambda variant) from Chile. These variants are currently monitored and tracked globally, and classified to be Variants of Concern (VoC) and Variants of Interest (VoI) by the WHO.

In this present study, we evaluated different SARS-CoV-2 sequences bioinformatically through the use of FASTA parsing and alignment sequencing to gain insight on the rate of substitution in different variants of the virus. Our main goal was to estimate the rate of substitution in SARS-CoV-2 and understand new arising variants and their spread and to see if the rate differed between each variant. We also calculated the sequence divergence between different SARS-CoV-2 variants which shows us that the virus is in fact mutating at an alarming rate in a short time compared to the original strain (Wuhan-Hu-1).

Methods

For this study, we chose to use the NCBI Virus database to pick our sequence data because we could easily download accessions in the form of FASTA files. Once we had the FASTA files from eight of the variants of interest, we combined them into one FASTA file. The file was then put into a Multiple Sequence Alignment software called MUSCLE. This returned a FASTA file with aligned versions of each sequence.

We used python code to parse this file and calculate the sequence divergence rates and the substitution rate. To calculate the sequence divergence percentage, each variant sequence was compared with each other to find the nucleotide positions that did not match, then divided the number of differences by the length of the shortest sequence.

We calculated the substitution rate under the Jukes & Cantor (JC69) model of substitution, using the equation (Dalhousie University, 2017):

$$\mu = -\frac{3}{4t} ln(1 - \frac{4}{3}p)$$

where,

μ = mutation rate
t = time
p = probability/divergence

Each variant was calculated using the divergence value from the original virus, and time was measured as the number of days from the original virus to the first recorded date of the variant.

To visualize the data, we created tables using the Pandas software package and graphed the data using the Matplotlib and Seaborn software packages.

A copy of our python code can be found in a Jupyter Notebook referenced in the supplementary data.

Results

Multiple Sequence Alignment (MSA)

Eight nucleotide sequences and one reference sequence of SARS-CoV-2 variants were collected from the NCBI database. They were parsed as a FASTA file, and aligned using the MUltiple Sequence Comparison by Log-Expectation (MUSCLE) program. The output file that was obtained from MUSCLE was put in an alignment visualization software called MView.

The different sequences are not of the same length at first, due to the different mutations. In order to find the positions of mutations, gaps were added during the alignment process, so that the characters in the same position correspond to the character at the original position. Doing alignments is a very useful step for studying mutations, or when you want to compute sequence divergence of different sequences.

The file is available in Supplementary Data.

Sequence Divergence

Sequence Divergence is the percent difference in nucleotide sequence between two related DNA sequences. A heatmap (Figure. 1) was generated, showing the pairwise sequence divergence between all the 9 variants. It was calculated by comparing every nucleotide sequence to each other. As seen in Figure.1, every sequence has a divergence, which means that mutations are present in all of them. From our data, we

found that the Eta variant has the highest percent divergence, compared to other variants. Data for sequence divergence can be found in table 1.

Rate of Substitution

Substitution Rate is the rate of nucleotide substitutions in a genome sequence over time, also known as the mutation rate. We calculated the rate of substitution under the Jukes & Cantor (JC69) model of substitution.

The substitution rate of each variant was calculated against the original SARS-CoV-2 virus where time is measured in days from the date of the first recorded sample of the original virus to the first recorded sample of the variant. The results of these calculations are shown in figure 2 where Alpha has the highest rate of substitution and Delta has the lowest. Data for the substitution rate can be found in table 2.

The figures and data we obtained gave us a pretty good understanding about the trends of the mutations and substitution rate of SARS-CoV-2 virus. It is also clear that the virus is still continuously mutating, but only further research can tell if these mutations have detrimental effects on the human body. We did this study with a lot of limitations, which will be discussed in the discussion section. Therefore, more data is required to draw better conclusions.

Discussion

Sars-CoV-2 is a constantly evolving new virus that researchers have been studying for less than two years. With the predicted trajectory of the virus, we were trying to find the substitution rate of the virus to better understand the evolution of the mutations. By calculating the sequence identity, days since the original virus emerged, and then the rate of substitution of each variant of interest, we were able to find how the rate diverged from each other. Moving forward, this data will have an impact on how researchers look at other variant's substitution rates so that they will be able to make predictions about the virus in the future.

Our approach to finding the substitution rate was to calculate the sequence identity divergence of the eight variants and divide that by the number of days it took the variants to mutation to find the rate of substitution for each variant. What we found was that each variant had a different substitution rate. What we found was that each variant had a different substitution rate. Other studies concluded that the rate of substitution is similar to the rate of transmission. We tested a similar correlation between the rate of substitution but for different lineages of the virus not specific to a location (Chen et al.,

2020). Another way of tracking the evolution of the virus is by finding the rate of substitution of amino acid sequences instead of using nucleotide sequences (González-Candelas et al., 2021).

A limitation of this study was that the earliest date for each lineage was stated differently on each database. To calculate the rate of substitution, we used the data from cov-lineages.org and used the same website for all the variant dates. Another limitation was how small the variant sample size was. We only used one accession for each variant calculation instead of a random sample of a higher number. We also did not specify the location of each sample. It is possible that the rate of substitution varies geographically. Using a larger sample size of varying geographical locations could have reflected an average substitution rate globally.

The findings of the substitution rate prompt additional hypotheses and questions about the virus and its variants. Our calculations suggest that the same methods should be done with multiple accessions of the same variant or collected in a specified location. Future studies based on our data could predict the trajectory of the virus and the future evolution involving the mutations. This data could be correlated to the rate at which people need to be vaccinated and useful for future Sars-CoV viruses.

In the short amount of time researchers have known about Sars-CoV-2, the knowledge of the substitution rate helps understand the evolution of the virus and how we can counteract its effects from a public health perspective. This information will be useful with vaccine development and production and will be helpful in the preventative measures towards epidemics and pandemics.

Figures

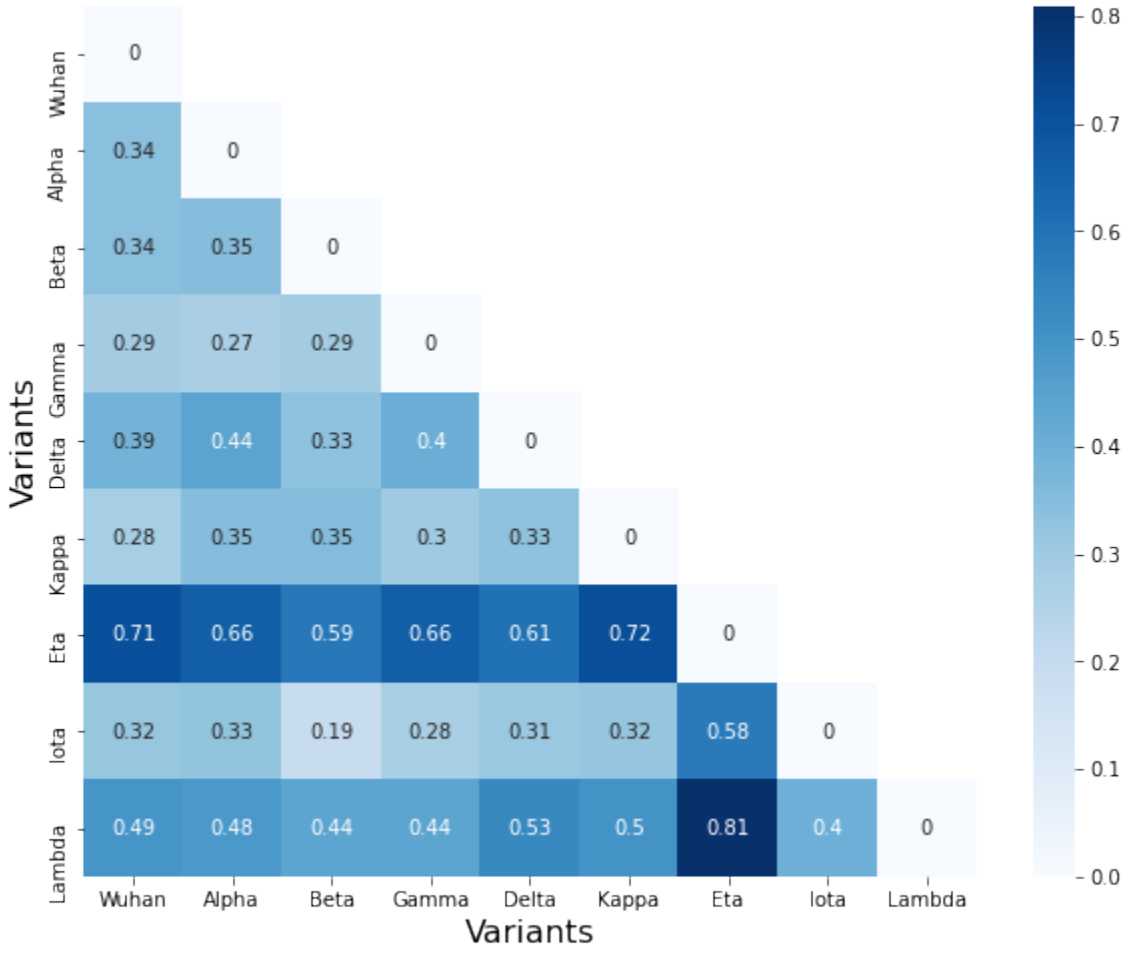

Figure 1: Sequence divergence of all variants of SARS-CoV-2 virus. Calculated as the percent difference in nucleotide sequence between two related DNA sequences. The divergence of each variant was calculated against every other variant by taking the number of nucleotides that were different in each sequence and dividing that by the length of the smaller sequence. Values closer to zero are lighter in color and represent a smaller percentage of different nucleotides, and higher values and darker blue represent a larger percentage of different nucleotides.

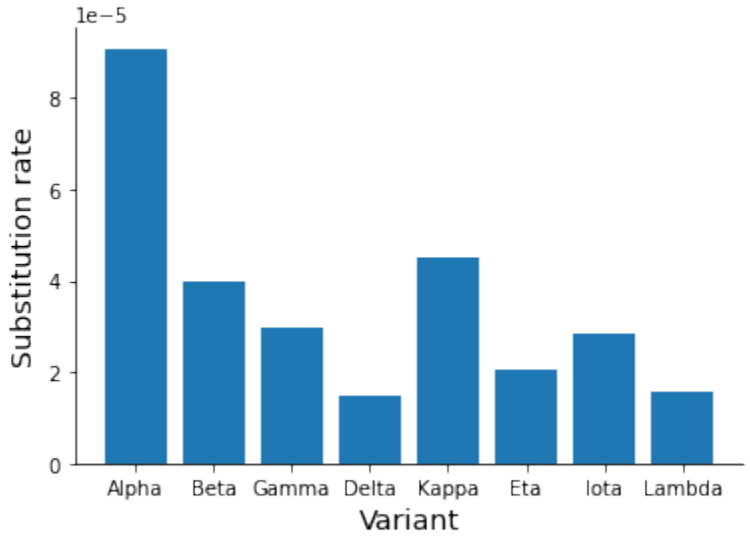

Figure 2. SARS-CoV-2 variant rate of substitution. Calculated under the JC69 model substitution rate. Time is measured in days from the first recorded sample of the original SARS-CoV-2 virus on December 31, 2019. Larger numbers represent more mutations in a faster time frame.

Tables

Table 1. Sequence divergence values in percent.

Variant	Wuhan	Alpha	Beta	Gamma	Delta	Kappa	Eta	Iota	Lambda
Wuhan	0.000000	0.344378	0.344378	0.290882	0.387843	0.284195	0.708817	0.317630	0.494834
Alpha	0.344378	0.000000	0.347721	0.274165	0.444682	0.347721	0.662008	0.327661	0.481460
Beta	0.344378	0.347721	0.000000	0.290882	0.334348	0.351065	0.585108	0.193922	0.437995
Gamma	0.290882	0.274165	0.290882	0.000000	0.404561	0.300913	0.662008	0.277508	0.437995
Delta	0.387843	0.444682	0.334348	0.404561	0.000000	0.334348	0.605169	0.310943	0.531613
Kappa	0.284195	0.347721	0.351065	0.300913	0.334348	0.000000	0.715504	0.320974	0.501521
Eta	0.708817	0.662008	0.585108	0.662008	0.605169	0.715504	0.000000	0.578421	0.809121
Iota	0.317630	0.327661	0.193922	0.277508	0.310943	0.320974	0.578421	0.000000	0.401217
Lambda	0.494834	0.481460	0.437995	0.437995	0.531613	0.501521	0.809121	0.401217	0.000000

Table 2. Substitution rate values.

Alpha	Beta	Gamma	Delta	Kappa	Eta	Iota	Lambda
9.08E-05	3.97E-05	2.97E-05	1.46E-05	4.52E-05	2.06E-05	2.84E-05	1.59E-05

Supplementary Data

Project code jupyter notebook: project2.ipynb

Variant FASTA files: AlphaVariant.fasta, BetaVariant.fasta, DeltaVariant.fasta, EpsilonVariant.fasta, EpsilonVariant.fasta, EtaVariant.fasta, GammaVariant.fasta, IotaVariant.fasta, KappaVariant.fasta, LambdaVariant.fasta, Wuhan-Hu_1.fasta

Multiple Sequence Aligned FASTA file: covid_variants_output1.fasta

Sequence divergence table: divergence_values.csv

Substitution rate table: sub_rate_values.csv

Multiple Sequence Alignment (MSA) Visualization:
https://www.ebi.ac.uk/Tools/services/rest/mview/result/mview-I20210818-070700-0353-15360301-p1m/aln-html

The MUSCLE software, source code and test data are freely available at: http://www.drive5.com/muscle.

NCBI SARS-CoV-2 Data Hub:
https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049

GISAID: https://www.gisaid.org

Sources

Chen, Yi-Hau, & Wang, Hsiuying. (2020). Exploring Diversity of COVID-19 Based on Substitution Distance. Infection and Drug Resistance, 13, 3887-3894.

Cov-lineages.org. Cov. (n.d.). https://cov-lineages.org/.

Dalhousie University. (2017). Neutral theory Topic 4: Estimating the rate of substitution. http://awarnach.mathstat.dal.ca/~joeb/biol3046/PDFs/Supp/Supp3a_EstRates.pdf.

González-Candelas, Fernando, Shaw, Marie-Anne, Phan, Tung, Kulkarni-Kale, Urmila, Paraskevis, Dimitrios, Luciani, Fabio, . . . Sironi, Manuela. (2021). One year into the pandemic: Short-term evolution of SARS-CoV-2 and emergence of new lineages. Infection, Genetics and Evolution, 92, 104869.

Maio, N. D., Walker, C. R., Turakhia, Y., Lanfear, R., Corbett-Detig, R., & Goldman, N. (2021). Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. Genome Biology & Evolution, 13(5), 1–14. https://doi-org.offcampus.lib.washington.edu/10.1093/gbe/evab087

Spratt, A. N., Kannan, S. R., Woods, L. T., Weisman, G. A., Quinn, T. P., Lorson, C. L., Sönnerborg, A., Byrareddy, S. N., & Singh, K. (2021). Evolution, correlation, structural impact and dynamics of emerging SARS-CoV-2 variants. Computational and Structural Biotechnology Journal, 19, 3799–3809.

WHO. Tracking SARS-CoV-2 variants https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/