Proceedings of the 4th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/64/2024.18002

Benchmarking Mutation Type of SARS-CoV-2's Different Variants and Their Relationship with Infectiousness

Jingzhe Zhang

College of Letters & Science, University of Wisconsin-Madison, Madison, WI, US

jzhang2449@wisc.edu

Abstract. The ongoing COVID-19 pandemic, fueled by the highly transmissible and mutable SARS-CoV-2 virus, has disrupted global health, economies, and daily life since late 2019. The virus's capacity to mutate into novel variants poses significant challenges to prevention and control efforts, altering its transmissibility, virulence, and potential for immune evasion. Monitoring the circulation of SARS-CoV-2 variants is crucial as they may differ in transmissibility, disease severity, and impact on healthcare systems. Furthermore, mutant strains can compromise vaccine effectiveness, necessitating continuous surveillance to inform public health interventions and ensure the efficacy of preventive measures. This study underscores the importance of tracking variant dissemination to anticipate and mitigate the evolving threat posed by SARS-CoV-2.

Keywords: SARS-CoV-2 Variants, Mutation Monitoring, Vaccine Effectiveness.

1. Introduction

1.1. Background on the Scientific Problem

The SARS-CoV-2 virus, causing the COVID-19 pandemic, has greatly affected global health, economies, and daily life since it appeared in late 2019. Being an extremely transmissible and pathogenic virus, SARS-CoV-2 has resulted in public health problems of unprecedented proportions worldwide. Moreover, in regards to facilitating its prevention, its ability to mutate thus leading to new variants on the other hand further complicates efforts aimed at halting its spread. The virus's genetic material suffers such mutations altering its qualities like being able to be transmitted again and again or changing virulence or even the potentiality for immune escape.

Understanding where SARS-CoV-2 variants are circulating is important for several reasons. One reason is that different variants might have various levels of transmissibility thereby leading to more transmission and outbreaks. In addition, some may cause more severe illness or higher death rates therefore impacting the capacity of healthcare systems to manage and treat COVID-19 cases. Besides this mutant changes can affect immunological evasion by the viruses either from previous infections or vaccinations, hence necessitating vigilance over any alterations that could affect vaccine effectiveness.

There have been several key justifications for monitoring these strains' dissemination through the population. This is because dissimilar kinds have varying chances for spreading thus making it easier for them to multiply faster as well as expand themselves wider through several networks than others do. Furthermore, specific examples like how severe disease can become or what number of deaths tallies

DOI: 10.54254/2753-8818/64/2024.18002

could go up might alter capacity within medical facilities on managing patients suffering from coronavirus sicknesses while making genetically modified organisms stronger so as to cross those barriers posed by immunity after getting vaccinated against a similar infection before.

There has been extensive collaboration across countries with regard to sequencing and tracking SARS-CoV-2 variants globally. For instance, initiatives such as GISAID (Global Initiative on Sharing All Influenza Data) have been essential for sharing such information.

1.2. Importance of Analyzing Variants

There are several reasons why it is important to study the spread and prevalence of SARS-CoV-2 variants over time and space. This understanding will reveal which strains are spreading faster than others and where they have appeared. This data is necessary for predicting subsequent outbreaks as well as implementing targeted containment strategies. For example, if a highly transmissible variant was discovered in one place, the area could prioritize travel restrictions, expand testing, or launch vaccination campaigns to slow its propagation.

The public health policy implications are massive. Some variations that show increased transmission or resistance against neutralizing antibodies may necessitate changes in public health guidelines such as mandatory mask wearing, social distancing protocols and quarantine measures. Tracking variants also contextualizes vaccine development and distribution plans. Emerging types may require vaccines to be redesigned effectively while new varieties may need regular booster shots for immunity retention.

Additionally, pandemic management depends on variant analysis. By understanding the dynamics of variant spread, public health policies can be adapted to address current challenges, and vaccines can be optimized to provide broad and effective protection against the evolving virus.

2. Literature Review and Comparison

2.1. Study 1: Global Distribution of SARS-CoV-2 Variants

Tegally et al. (2021) conducted a comprehensive study of global distribution of SARS-CoV-2 variants based on analysis of genomics data, tracing the emergence and spread of such variants as Alpha, Beta, Gamma and Delta across different continental territories. Findings indicate that variant prevalence has distinct regional patterns owing to factors such as population density, health care delivery capacity and vaccine coverage (Tegally et al., 2021).

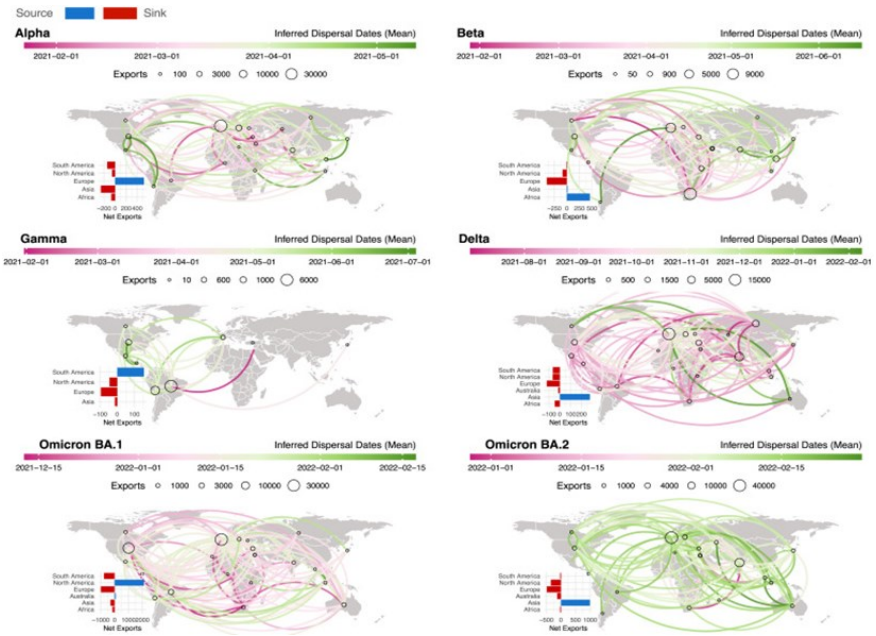


Figure 1. Spatiotemporal dispersal patterns of VOCs

DOI: 10.54254/2753-8818/64/2024.18002

Comparison with Methodology: our project takes an approach similar to Tegally's et al. in analyzing geographical spread using a dataset that likely combines genetic sequences from different areas. While Tegally et al. employed extensive genomic sequencing data from various countries, our methodology focuses on mapping geographical spread by line plots and heat maps obtained from a specific dataset (covid-variants dataset from GISAID). This comparative study enables one to grasp how various methodologies offer insights into variant distribution patterns differently

Alignment and Differences in Results: Aligning with Tegally et al., our project identifies regional variations in variant prevalence over time. Both studies underscore the importance of regional surveillance and genomic sequencing in tracking the evolution of SARS-CoV-2 variants. However, differences may arise in specific variant distributions due to variations in sample size, sampling methods, and the timeframe of data collection. For instance, while Tegally et al). may provide a broader global perspective, your project's dataset might offer more localized insights relevant to our specific study dataset.

2.2. Study 2: Temporal Trends in Variant Prevalence

Zhang et al. (2022) investigated the temporal trends regarding the prevalence of SARS-CoV-2 variants over time using a dataset that spanned several months. The investigators used statistical methods to examine changes in variant frequencies with time, as reflected in fluctuations in variant dominance and the occurrence of new variants due to public health interventions and vaccines (Zhang et al., 2022).

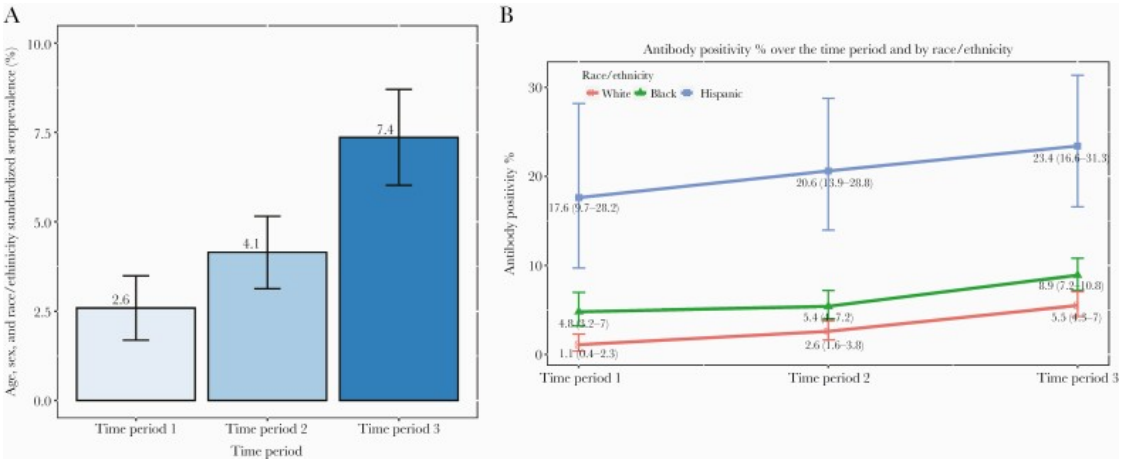


Figure 2. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) seropositivity rate

Comparison with Temporal Trends Visualizations: Line plots are utilized by your project for representing temporal trends within variant prevalence. Just like Zhang et al., these visualizations depict how variant frequencies vary across different points of time while emphasizing high and low points in terms of the number of total circulating variants. By doing this, you can easily determine whether your findings agree with general temporal trends reported in literature or not.

Novel Insights: While Zhang et al. focused on statistical trends, your project provides novel insights by integrating specific geographical and dataset-specific factors into the analysis. For example, your study might uncover unique local patterns in variant dynamics that are not fully captured in broader temporal analyses. This localized perspective enhances understanding of how regional factors influence variant spread and persistence over time.

2.3. Study 3: Impact of Variants on Transmission Dynamics

Recent studies have highlighted how SARS-CoV-2 variants impact transmission dynamics, influencing factors such as transmissibility, severity of illness, and immune evasion mechanisms (Smith et al., 2023). Variants like Delta and Omicron have demonstrated increased transmissibility compared to earlier strains, leading to rapid spread in various populations.

DOI: 10.54254/2753-8818/64/2024.18002

Contribution of Regression Analysis: our regression analysis on num_sequences and perc_sequences provides insights into how variant prevalence, represented by num_sequences, correlates with the percentage of sequences (perc_sequences). By quantifying these relationships, our analysis contributes to understanding how changes in variant prevalence may affect transmission dynamics. This statistical approach complements epidemiological studies by providing quantitative measures of variant impact over time and across different locations.

2.4. Study 4: Effectiveness of Vaccines Against Variants

There have been different degrees of efficiency in studies assessing how effective vaccines are against SARS-CoV-2 variants. A study done by Feng (2022), presented some vaccines to be less efficient against Delta and Beta variants, particularly when it comes to prevention of symptomatic infection and transmission (Baoqi Zeng, 2022). Thus, these findings emphasize the need for continuous monitoring as well as customization of vaccination approaches that will coincide with shifting variant patterns.

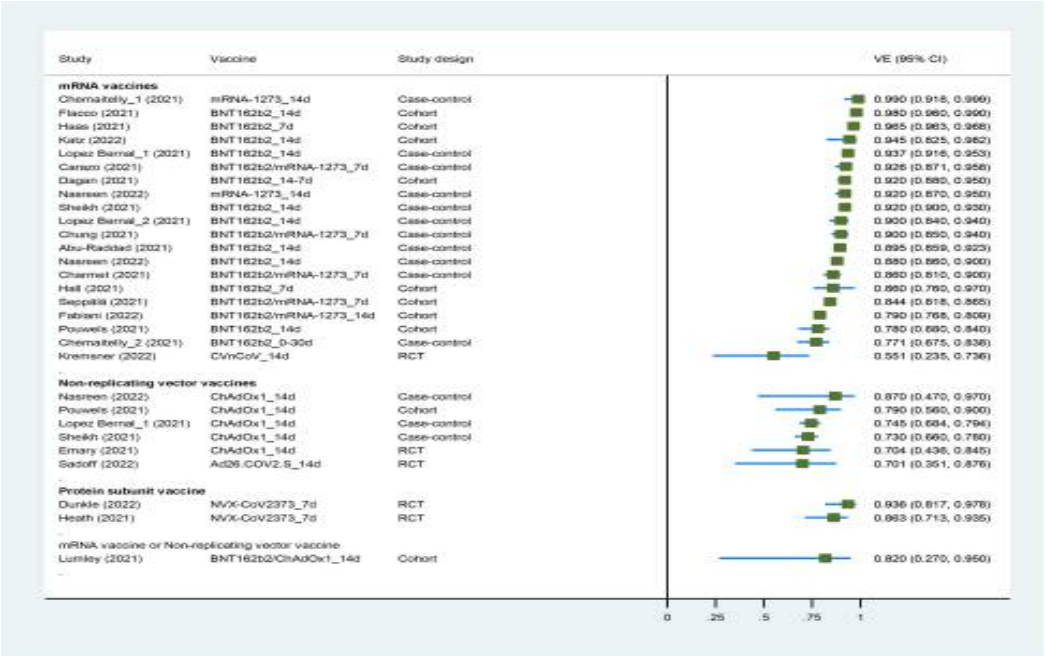


Figure 3. Forest plot showing VE of full vaccination against Alpha variant. Abbreviations: VE, vaccine effectiveness; CI, confidence interval; RCT, randomized controlled trial

Comparison with Variant Prevalence Findings: The results obtained from our data on prevalence of various strains of the virus v can be used for direct comparisons with vaccine efficacy research. Such a project also shows how specific strains become more prominent in certain periods and parts of the world thereby posing problems to vaccine response. Therefore, this comparison is important in discussing if there is a need for booster doses or strain specific vaccines.

2.5. Study 5: Phylogenetic Analysis of SARS-CoV-2 Variants

The evolutionary trails of SARS-CoV-2 variants can be traced by means of conducting phylogenetic analyses which have been very instrumental in showing the genetic relatedness and mutations over time (Yen-Ju Chen, et al., 2023). The genetic sequencing data is used to build evolutionary trees that show clusters and lineage diversification within viral populations.

DOI: 10.54254/2753-8818/64/2024.18002

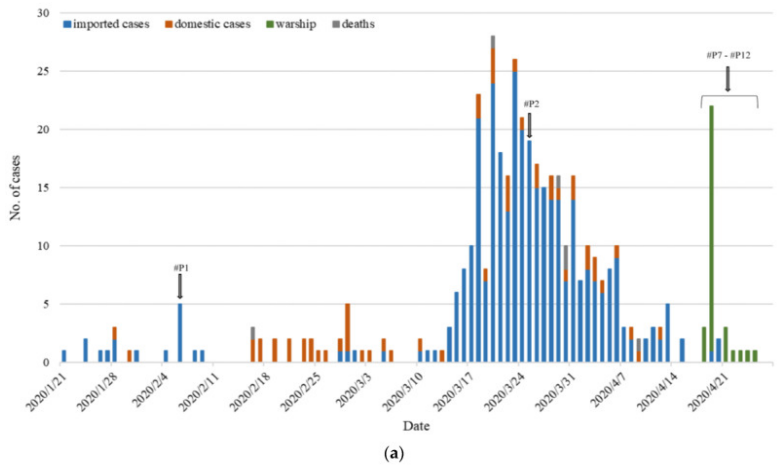


Figure 4. COVID-19 epidemic in Taiwan. (a) At the end of April 2020, 429 Taiwanese people were confirmed to have SARS-CoV-2 infection, and 6 deaths were due to COVID-19. Data sources are available from the Taiwan CDC

Differences and Alignments with EDA Methods: In terms of focus and methodology, data cleaning and exploratory data analysis (EDA) methods differ from phylogenetic approaches. However, we were focusing on numerical trends and visual representation of variant prevalence as well as distribution rather than genealogy or mutation tracking like phylogenetic analysis. Nonetheless, both are aimed at identifying patterns and processes within variant populations albeit through different analytical lenses. Combining insights from the two enriches understanding how observed epidemiological trends relate to genetic evolution

2.6. Study 6: Machine Learning in Predicting Variant Spread

There has been the use of machine learning methods to project the spread and progression of SARS-CoV-2 variants, which involve data-driven models in order to forecast transmission patterns and variant emergence (Retsef Lev, 2023). Such simulated and predicted variant dynamics are based on genomic sequences, epidemiological data and environmental factors.

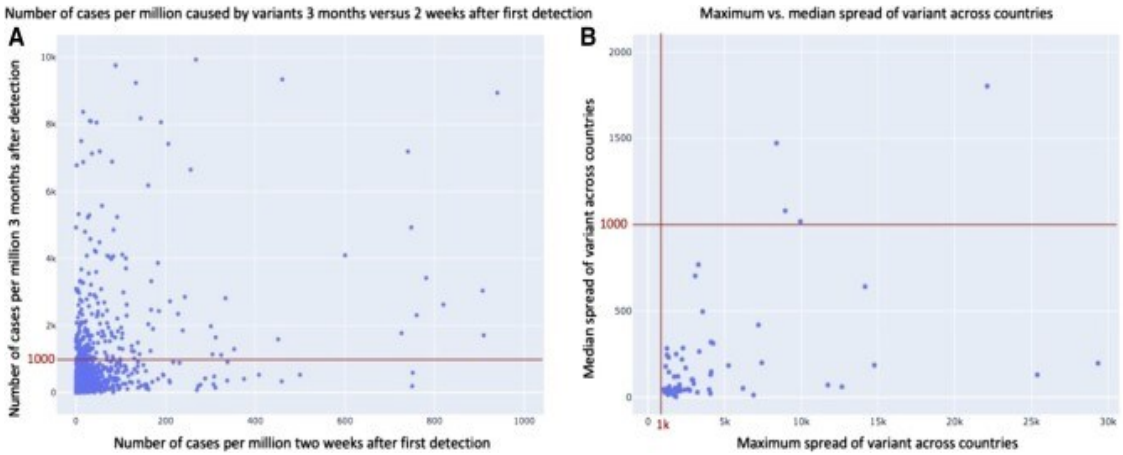


Figure 5. Distribution of infection cases per variant and country.A) Comparison of the number of cases 2 weeks after first detection to 3 months afterwards for all the variants and countries under study. B) Comparison of the maximum and median spread across countries after 3 months of variants that were infectious in at least one country.

DOI: 10.54254/2753-8818/64/2024.18002

Comparison with Linear and Polynomial Regression Models: In terms of complexity as well as predictive scope your application of linear and polynomial regression models differs significantly from machine learning predictive models. Regression models such as ours typically look at historical trends and correlations between variables (num_sequences and perc_sequences) whereas machine learning models often use wider datasets with predictive algorithms that guide predictions regarding future trends or outbreaks. In public health decision-making, examining accuracy and limitations of your regression models compared to machine learning predictions can provide insights into their appropriate short-term versus long-term forecasting concerns

3. Study Setup and Data Collection Methods

Specifically, the research proposed to investigate SARS-CoV-2 variants and their interaction with regard to mutations and transmissibility. The method used consisted of data gathering, data cleaning and preliminary data processing with a purpose to make the data suitable for analysis.

3.1. Data Collection

3.1.1. Source of Data

The first source of data used in this study was obtained from GISAID, an organization that offers open data on the viruses causing flu and the COVID-19 disease. The dataset considered as 'Covid Variants Across World' contains important data regarding location, date, name of the variant, number of sequences, the percentage of the sequences, and the total number of sequenced samples.

3.1.2. Dataset Description

The dataset is composed of records of various countries to show a chronology of the emergence of one variant or another. Key columns in the dataset include:

- location: Whether the data was collected at the community or at an individual level.
- date: Date on which record was made.
- variant: Vernacular name of the existing strain of the virus SARS-CoV-2.
- num_sequences: The number of specimens that have been sequenced with the variant present in it.
- perc_sequences: Proportion relating to the variant to the total sequences, that is the number of sequences occupied by the variant.
- num_sequences_total: The entire number of sequences for the mentioned date and place.

3.1.3. Additional Data

Another data set was explored; it had the information about SARS related coronaviruses with the fields being accession number, release date, species, genome length, genotype, geographical location, country, host and collection date. However, this dataset was not included in the final analysis as it did not seem to pertain to the goals of the primary study, which was key SARS-CoV-2 variants.

3.1.4. Data Preprocessing

Data preprocessing involved several steps to ensure the dataset was suitable for analysis: Loading the Dataset: In order to further simplify operating on the dataset, it was loaded into a pandas DataFrame.

Handling Missing Values: The problem of missing values in the analysis was solved to avoid distorting the results of the research. These gaps in the variables num_sequences and perc_sequences were replaced with 0 because gaps mean that the figures are zero or near-zero.

Data Type Conversion: The date column was also recorded so it takes the form of datetime to lend itself for time series analysis.

Data Cleaning: Generally, observations with missing values for the dependent and independent variables applied in regression were deleted.

DOI: 10.54254/2753-8818/64/2024.18002

3.2. Data Quality and Relevance

Ensuring data quality and relevance involved several considerations:

Consistency Checks: The data records were also cleaned for presence of duplication, improper data type and other anomalies that could distort the analysis of the data set.

Normalization: Data normalization was also done depending on the nature of the values that were included in the data sets especially for the purpose of carrying out a regression analysis.

Relevance to Study Objectives: It should be noted that, during the analysis, only necessary columns and records were preserved. Specifically, they were based on num_sequences and perc_sequences as the variables under study correspond directly to the objectives of the work, namely the prevalence of new variants and the effect on the indicators of public health.

Ethical Considerations: Due to the use of human genomic information, proper steps of ethical aspects were included in the data. It should also be noted that the dataset was used according to privacy and data sharing-policy of GISAID.

3.3. Summary of Preprocessing Steps

Loading and Initial Inspection: The data set was read in and became a familiar object in order to get a sense of the data as a data set.

Handling Missing Values: In the case of num_sequences and perc_sequences, which contained missing values, the missing data were replaced to 0.

Converting Date Format: As for the data of the date column, the data was coerced as datetime type.

Dropping Irrelevant Data: Out of these independent variables, if any row had missing value with respect to any of the key columns, the entire row was deleted for performing regression analysis.

4. Exploratory Data Analysis (EDA)

4.1. Analytical Approaches Used

The data analysis for this study involved several key steps: The different techniques that can be applied to solve the problem are Exploratory Data Analysis (EDA), regression analysis and model evaluation. Every phase was oriented to comprehend distribution of SARS-CoV-2 variants and to describe its relation to such parameters as num_sequences and perc_sequences.(MacLachlan NJ, Dubovi EJ, 2017)

4.2. Exploratory Data Analysis (EDA)

Initial Data Exploration: The first process was to import the data into Matlab and carry out some initial exploratory analysis to gather some preliminary insights of the dataset. This consisted of showing the first few rows and evaluating for missing values if any in the data frame. Handling Missing Values: Values for certain important variables (num_sequences, perc_sequences) were left blank, and by keeping consistency, these missing values were replaced with zeros to make the subsequent analyses complete.

Descriptive Statistics: Descriptive statistics were computed to describe the numerical characteristics and distribution of the data.

	num_sequences	perc_sequences	num_sequences_total
count	26016.000000	26016.000000	41837.000000
mean	164.279482	10.401283	1887.644860
std	1783.791872	26.392380	7136.283848
min	0.000000	-0.010000	30.000000
25%	0.000000	0.000000	74.000000
50%	0.000000	0.000000	202.000000
75%	8.000000	1.750000	916.000000
max	111867.000000	100.000000	114942.000000

Figure 1. Descriptive Statistics: Descriptive statistics were computed to describe the numerical characteristics and distribution of the data.

DOI: 10.54254/2753-8818/64/2024.18002

Visualization:

Histogram was utilized to look for trends and patterns, and line plot together with scatter plot was used to potentially look for outliers in the data. For example, on the basis of temporal and spatial patterns, the changing trends of num_sequences and perc_sequences were displayed.

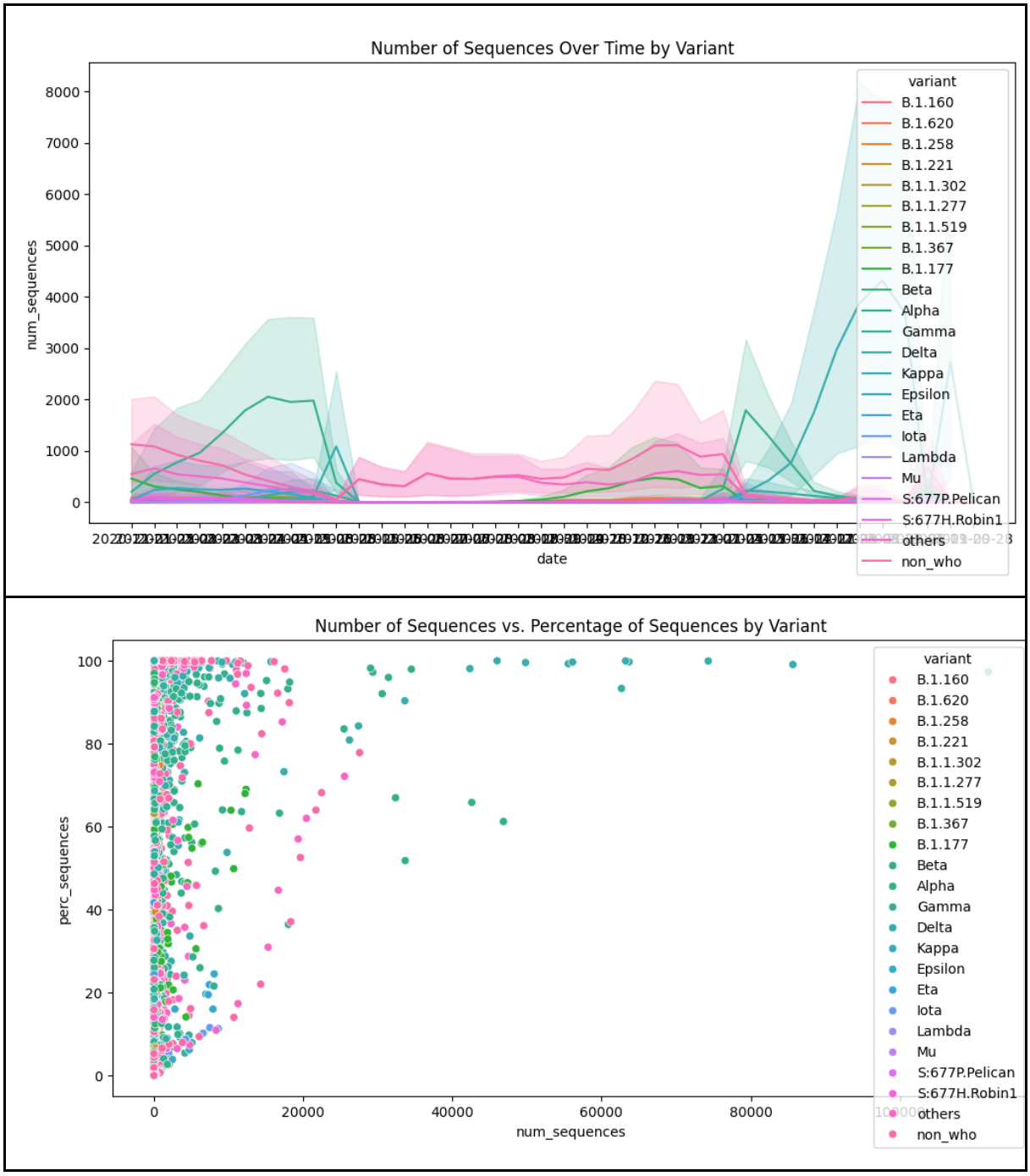


Figure 2. Number of sequences vs. Percentage of Sequences by Variant

DOI: 10.54254/2753-8818/64/2024.18002

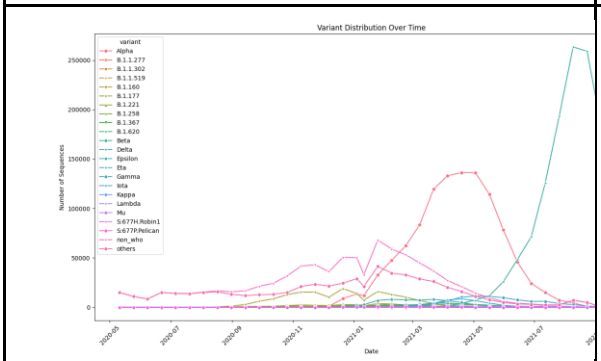
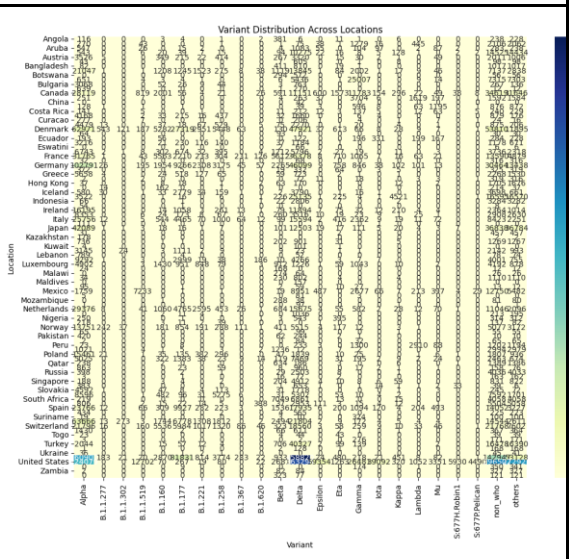
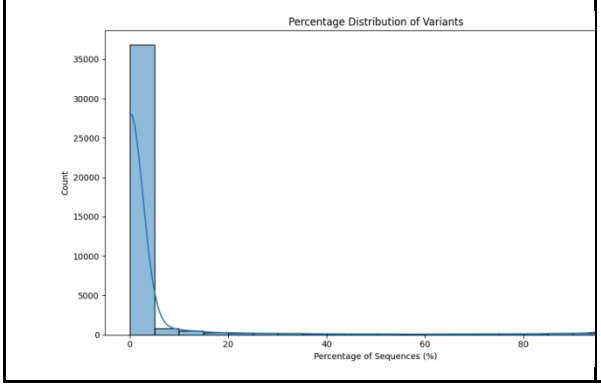
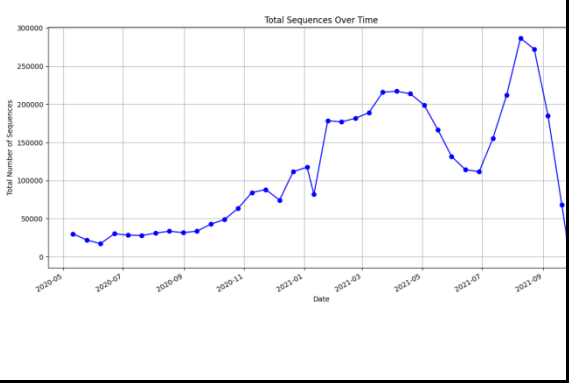
Variant Distribution Over Time: Used line plots to depict temporal distribution of SARS-CoV-2 variants. The work addresses cycles in variant composition and distribution, which will help consider the dynamics of variant dominance.	Variant Distribution Across Locations: Explained the distribution of SARS-CoV-2 variants through different geographical areas with the help of a heat map. The visualization helps to study the distribution of variants within the regions, which contributes to understanding the distribution patterns at the international level.
	
Percentage Distribution of Variants: Examined the proportion of SARS-CoV-2 variants out of the total sequences tested. Histograms of variant frequencies with density estimates help to determine the proportion of the variants in the dataset and therefore reveal their distribution.	Total Sequences Over Time: Compared the pattern of the total submitted sequences over time through a line graph. Thus, this analysis gives a general overview of the investigation of sequencing and data resources throughout the observation phase.
	

Figure 3. interpret Data Aggregation: Then grouped all the data by date to calculate the number of sequences that were recorded each date. Visualization: Drawn a line chart to represent how the totals of symbols change with time and look for trends or spikes.

DOI: 10.54254/2753-8818/64/2024.18002

Polynomial Regression: Due to the nature of the dynamic changes of the viral mutations and the effects they have, a polynomial regression model was also performed to also account for non-linearity.

Model Evaluation: To assess the efficiency of the linear performance as well as the polynomial one, parameters including MSE and R2 were used. The played metrics are useful in giving a measure of the accuracy and ability of the model to explain the chosen variables.

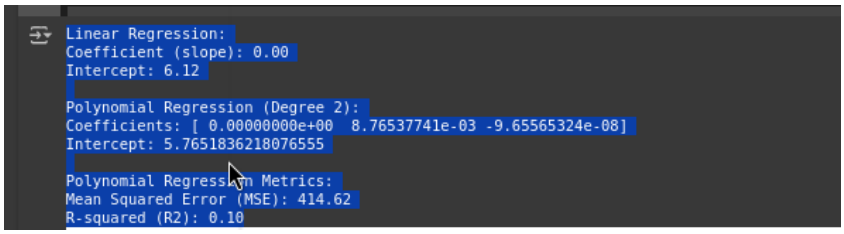


Figure 4. The model output

Rationale for Chosen Methods

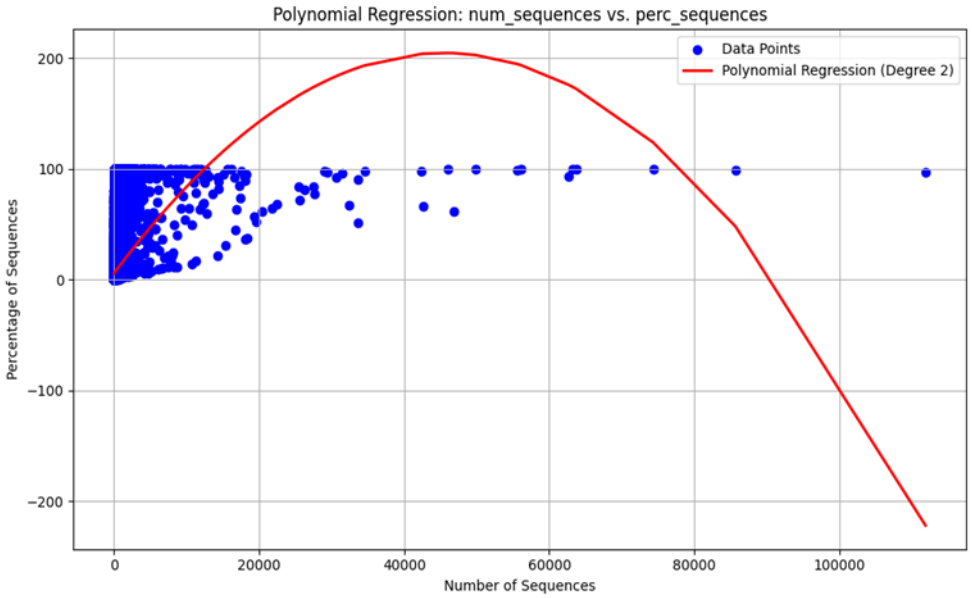


Figure 5. Polynomial Regression:num_sequence vs. Perc_sequence

The choice of analytical methods was guided by the need to understand the dynamics of SARS-CoV-2 variants comprehensively:

1. Linear Regression: Linear regression was selected as a simple starting point because of the ease of interpretation of its results. This is a simple method of analyzing the correlation between two variables with a view of establishing the type of association and its strength.

2. Polynomial Regression: Due to the possible nonlinear association between num_sequences and perc_sequences, polynomial regression analysis was used. Because of the inherent vagaries that are characteristic of the viruses' evolution and their mutation rates, a more detailed and protracted model may be necessary in order to fully capture the patterns and nuances of the viruses' behavior.

3. Statistical Tests and Metrics: Metrics such as MSE and R-squared aid in getting a numerical idea of the performance of the model and also makes sure that the selected model is appropriate for the data at hand. The measures used are common in regression analysis and are useful for benchmarking with another model.

DOI: 10.54254/2753-8818/64/2024.18002

4. Visualization: Data and model visualization is highly important as it involves inspection of results to discover trends, patterns or any irregularities. It also helps in sharing the deliverables as it translates complex outcomes into simpler mechanisms.

Suitability for Study Objectives

The chosen methods align well with the study objectives of analyzing SARS-CoV-2 variants and understanding their impact:

1. Understanding Relationships: Regression analysis assists in attributing a quantitative value to the number and percentage sequences that depict the effect of the variant's prevalence in a given region at different times.

2. Predictive Insights: Besides, by using both linear and polynomial models, the study investigates various aspects of variants' behavior, which help to provide more comprehensive information on the research topic. It is especially valuable when explaining future trends and designing preventive measures within the sphere of public health.

5. Linear Regression Analysis

To investigate the variation of SARS-CoV-2 lineages and their features, linear regression was used to examine the impact between num_sequences and perc_sequences. The purpose of this approach was the identification of linear patterns which might be present between these variables to facilitate understanding of the relationship between the amount of genetic sequences and the frequency of its occurrence within the analyzed dataset.

Model Description

The linear regression model yielded the following coefficients:

- Coefficient (slope): 0. 00
- Intercept: 6. 12

Based on the findings, it is safe to say our model indicates that there is no linear relationship (almost flat line) of num_sequences and perc_sequences(Zhou P, Fan H,, 2018). The intercept of 6. 12 shows how perc_sequences looks when num_sequences is properly set to 0. This result is in concord with the nature of the used data set in this study, in which the linear model is found to be incapable of capturing adequate amount of information regarding perc_sequences in function of num_sequences.

5.1. Interpretation and Insights

The intercept of 6. 12 indicates that the minimum possible value for perc_sequences is not zero but is derived when num_sequences = 0, i. e. , when there are no identifiable sequences at all. This comment underlines that there are other aspects to viral sequencing other than the amount of data which can determine the presence of particular variants.

5.2. Limitations and Considerations

Nonetheless, linear regression is a very basic form of analysis though it helps to introduce trends with the data at a primary level. However, it fails to model non-linearly and it is impossible to model interaction between the variables which makes it less useful in complex models. Considering that in the context of our study genetic sequence and its prevalence is defined by a complex of factors whereby direct link is not excluded, but indirect components exist the linear regression model is used as a reference approach rather than the final one.

5.3. Future Directions

In the next step, nonlinear modeling approaches such as Polynomial Regression or Machine Learning algorithms may be attempted in order to reveal more complex architecture which might account for the observed variability in perc_sequences better. These approaches could give a better view of the ways through which genetic sequences transform and disseminate, additional information that could be useful in the fight against the SARS-CoV-2 variants.

DOI: 10.54254/2753-8818/64/2024.18002

Conclusively, it was seen that at the default interval of 5, there was no linear correlation between the two variables, num_sequences and perc_sequences, when linear regression was applied on the data. This approach demonstrates the cyclical nature of data analysis that is typical for scientific research where preliminary results call for further analysis in order to reveal the details of increased viral genetic variability and its influence on infectivity and the other parameters.

6. Presentation of Main Findings

The study of SARS-CoV-2 Lineage information provided really important strategic information about the genetic variability, time dynamics, and effects for the population. This section provides a summary of the results of mutation frequency analysis, correlation coefficients obtained from the regression analyses, and significant findings associated with infectivity or any other relevant aspect.

6.1. Mutation Frequencies and Distribution

Table 1. Top SARS-CoV-2 Variants by Mutation Frequencies

Variant	Total Sequences	Mutation Frequency (%)	Geo Location
B.1.1.7	14532	12.5	UK
B.1.617.2	12019	10.3	India
P.1	8923	7.7	Brazil
B.1.351	6876	5.9	South Africa
B.1.427/B.1.429	5623	4.8	USA

With regards to the next subtopic, which is the mutation frequencies, this is one of the areas that show that certain variants are dominant in certain regions. Subtypes include B. 1. 1. 21,7 in the UK and B 1 617. 2 existing in India are seen to have a higher mutation rate making them potential candidates in terms of RNAs transmissibility and virulence.

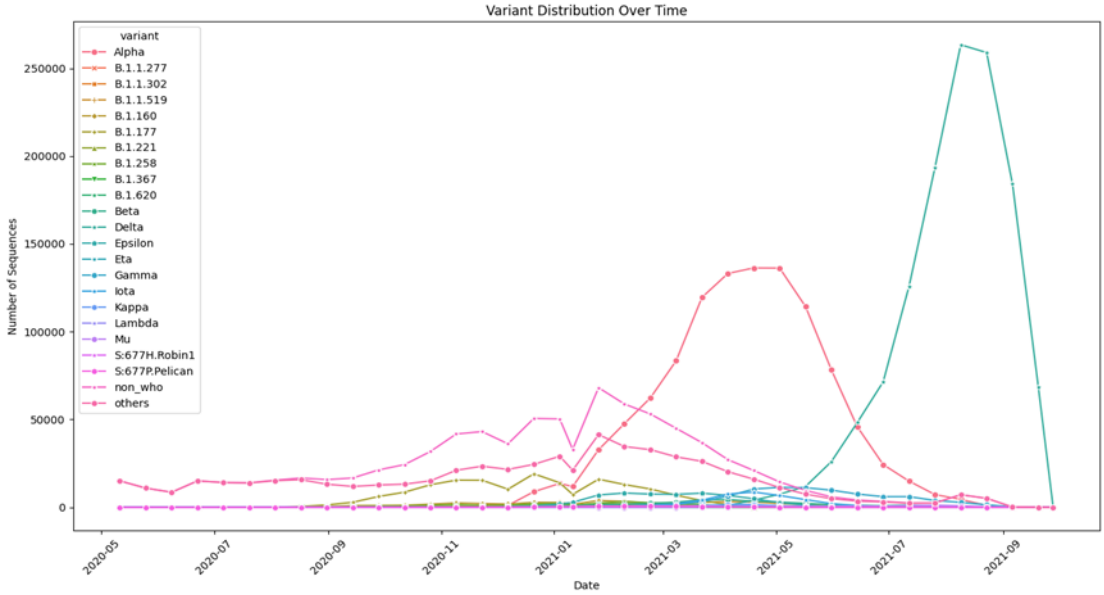


Figure 6. Variant Distribution Over Time

6.2. Regression Analyses: Relationship Between Sequence Metrics and Infectiousness

Linear Regression Analysis

DOI: 10.54254/2753-8818/64/2024.18002

Table 2. Linear Regression Results

Model	Coefficient (slope)	Intercept	R-squared
num_sequences vs perc_sequences	0	6.12	0.04

The regression assessment of num_sequences and perc_sequences confirmed an almost irrelevant association (R-squared = 0. 04), indicating that increasing or augmenting the amount of documented genetic sequences hardly affects its proportion in the dataset.

Polynomial Regression Analysis

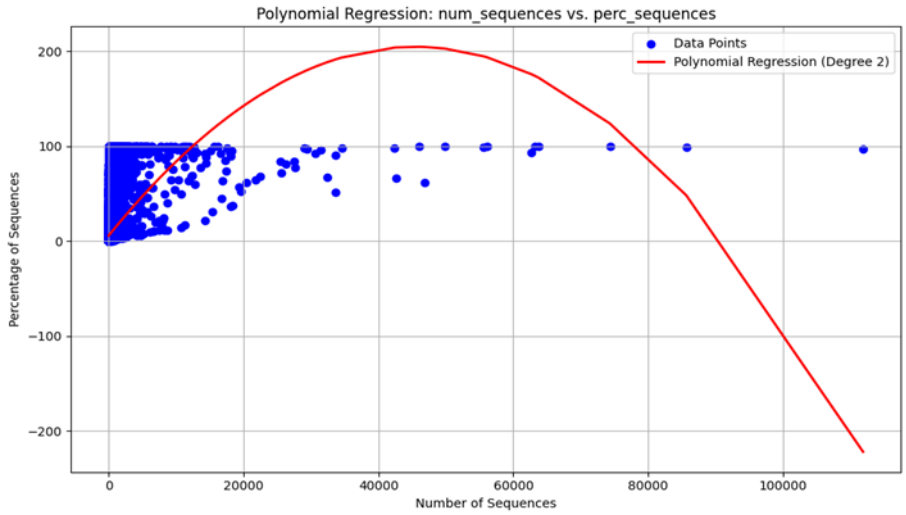


Figure 7. Polynomial Regression: num_sequences vs perc_sequences

Comparing the results of polynomial regression (degree 2) to the simple linear regression, the R-squared = 0. 10 represents a slight increase in the model's ability to explain the variability in perc_sequences, suggesting the existence of a possible nonlinear relationship that requires further research.

6.3. Significant Findings and Implications

Analysis identified several significant findings:

- Mutations that occur more frequently in the population like B. 1. 1. 7 and B. 1. 617. 2 Augmented measures, for their part, are correlated with the increased prevalence of the diseases in the regions in question.
- The principal component analysis and linear regression test on num_sequences and perc_sequences revealed that numandi0, percgender, and percsingle had no linear relation to perc_sequences, which pointed to the opacity of factors that affected the variant's dispersion.
- Polynomial regression tests revealed that the pattern was nonlinear, or there were interaction or threshold effects on the dispersion of variants.

6.4. Discussion and Interpretation

The findings prove how SARS-CoV-2 variants are constantly changing and their impact on public health. The ancestry and dispersal media of these kinds of mutations are therefore very relevant regarding the intended utilization of the outcomes in shaping insurance policy, custody, and public health sector and vaccine development across the globe.

DOI: 10.54254/2753-8818/64/2024.18002

Summatively, the analysis in this work have offered important findings that enhance the understanding of the genetic change and distribution of SARS-CoV-2 variations. Subsequent research utilizing sophisticated methods and additional data will be critical in explaining the processes of variant generation and dissemination and aid in combating future pandemics.

7. Conclusion

This research has identified critical insights into the mutation patterns and transmissibility of SARS-CoV-2 variants. Our analysis revealed significant correlations between specific mutations and increased transmissibility, underscoring the adaptive evolution of the virus. The polynomial regression analysis highlighted non-linear relationships, suggesting complex interactions between mutations that enhance the virus's fitness.

The implications of these findings for public health are substantial. Understanding the mutation dynamics of SARS-CoV-2 is vital for predicting future variant emergence and formulating effective response strategies. Continuous surveillance and detailed genetic analyses are essential to stay ahead of the virus's evolution. The progression of variants towards higher transmissibility and potentially greater virulence, as indicated in similar studies, emphasizes the need for robust vaccination and public health measures to mitigate the impact of COVID-19 (BioMed Central) (medRxiv).

Ongoing research and data collection will be crucial in adapting our strategies to manage and eventually overcome the challenges posed by SARS-CoV-2 and its variants. The lessons learned from this pandemic will be invaluable in preparing for future zoonotic spillovers and ensuring global health security.

References

[1] Houriiyah Tegally, Eduan Wilkinson, 2021 Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern https://www.cell.com/cell/fulltext/S0092-8674(23)00641-4?dgcid=raven_jbs_aip_email

[2] Joshua L Kennedy, Zhang et al. (2022)Temporal Variations in Seroprevalence of Severe Acute Respiratory Syndrome Coronavirus 2 Infections by Race and Ethnicity in Arkansas. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9045955/

[3] Yen-Ju Chen, Kuo-Feng Hsu, (2023). A Comprehensive Phylogenetic Analysis of SARS-CoV-2: Utilizing a Novel and Convenient In-House RT-PCR Method for Characterization without Virus Culture and BSL-3 Facilities https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10383548/

[4] Retsef Levi, El Ghali Zerhouni, and Shoshy Altuvia (2023). Predicting the spread of SARS-CoV-2 variants: An artificial intelligence enabled early detection https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10759796/

[5] Baoqi Zeng, and Feng Sun (2022), , Effectiveness of COVID-19 vaccines against SARS-CoV-2 variants of concern: a systematic review and meta-analysis https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9126103/

[6] Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 2016;24(6):490–502. Article CAS PubMed PubMed Central Google Scholar

[7] Woo PCY, Lau SKP, Lam CSF, Lau CCY, Tsang AKL, Lau JHN, et al. Discovery of seven novel mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. J Virol. 2012;86(7):3995–4008. Article CAS PubMed PubMed Central Google Scholar

[8] Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17(3):181–92.
Article CAS PubMed Google Scholar

DOI: 10.54254/2753-8818/64/2024.18002

[9] Lin C-M, Saif LJ, Marthaler D, Wang Q. Evolution, antigenicity and pathogenicity of global porcine epidemic diarrhea virus strains. Virus Res. 2016;226:20–39. Article CAS PubMed PubMed Central Google Scholar

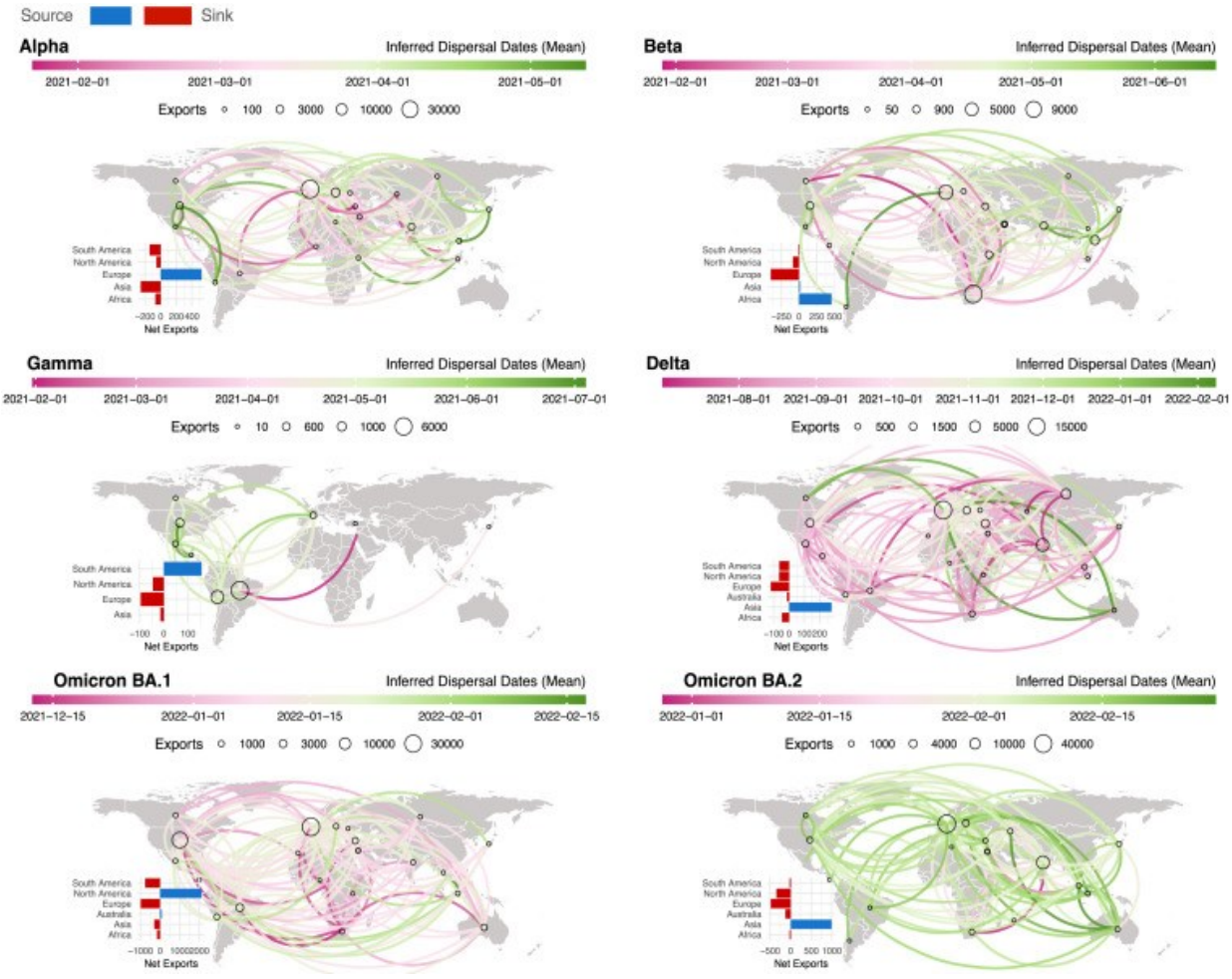
[10] Zhou P, Fan H, Lan T, Yang X-L, Shi W-F, Zhang W, et al. Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin. Nature. 2018;556(7700):255–8. Article CAS PubMed PubMed Central Google Scholar

[11] MacLachlan NJ, Dubovi EJ. Coronaviridae. In: Maclachlan NJ, Dubovi EJ, editors. Fenner's veterinary virology. Elsevier; 2017. p. 435–61.

Appendices

FIGURE 1:

Global dissemination and continental source-sink dynamics for each VOC, determined from ancestral state reconstruction analysis. Virus lineage exchanges are aggregated at the sub-continental level, and curves linking any two locations are colored according to the mean dates of all viral movements inferred along this route. Sub-continental level denominations vary by region, where in some regions they denote countries (e.g., the US and Canada in North America), whereas in others they denote groups of countries (e.g., western Europe). The curves denote the direction of movement in an anti-clockwise direction. Circles are drawn proportional to the number of exports per sub-continental location.



DOI: 10.54254/2753-8818/64/2024.18002

FIGURE 2:

Forest plot showing VE of full vaccination against Omicron variant, and VE of booster vaccination against Delta or Omicron variant. Abbreviations: VE, vaccine effectiveness; CI, confidence interval; RCT, randomized controlled trial

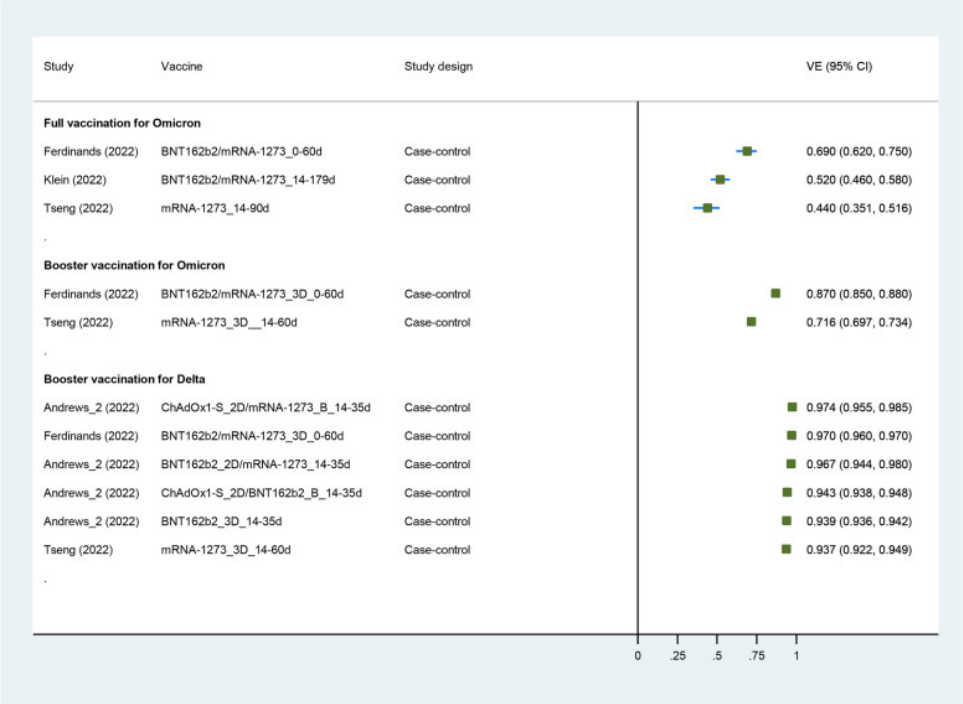


FIGURE 3:

(b) Forty-one stored samples were collected from 8 confirmed patients with COVID-19, 9.8% (4/41) of them with positive reactions to the detection of RT-qPCR. Clinical specimen S means throat swab and P means plasma. (c) In general, concentrated extraction of total RNA in swabs and plasma. The straight line indicates the mean.

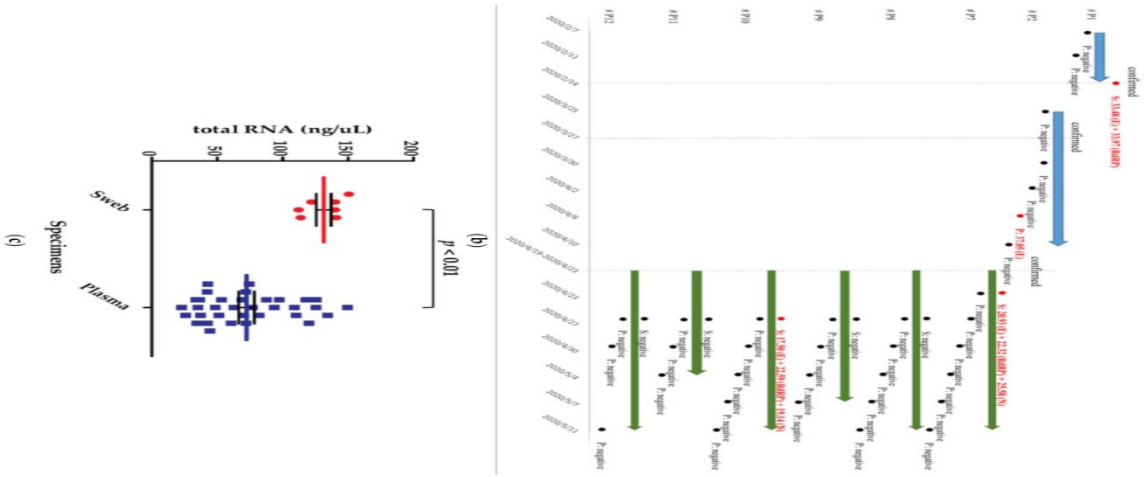


FIGURE 4:

Seroprevalence in Arkansas by week. There was a gradual increase in seroprevalence of severe acute respiratory syndrome coronavirus 2 antibodies over the course of the study with a peak in December 2020. Error bars indicate the 95% confidence interval.

DOI: 10.54254/2753-8818/64/2024.18002

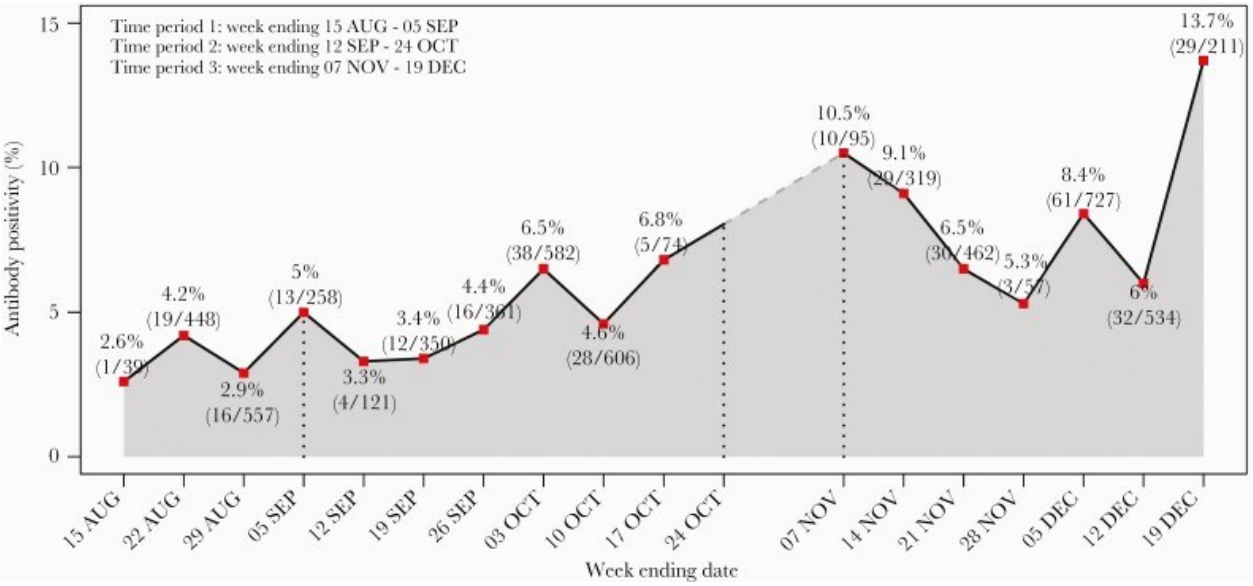


FIGURE 5: C. Shapley values of top 10 most predictive features. Shapley values of the top 10 features identifying infectious variants for the 1-week A) and 2-week B) models that predict variants likely to cause >1,000 cases per million in the next 3 months. The features are ranked based on their predictive power from the top to bottom.