Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

ViralVar: a web tool for multilevel visualization of SARS-CoV-2 genomes

Arghavan Alisoltani[1,2,3]*, Lukasz Jaroszewski[4], Adam Godzik[4], Arash Iranzadeh[5], Lacy M. Simons[2,3], Taylor Dean[2,3], Ramon Lorenzo-Redondo[2,3], Judd F. Hultquist[2,3], Egon A. Ozer[2,3]*

Affiliations

[1]Department of Microbiology-Immunology, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA

[2]Department of Medicine, Division of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA

[3]Center for Pathogen Genomics and Microbial Evolution, Havey Institute for Global Health, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA

[4]University of California Riverside School of Medicine, Biosciences Division, Riverside, California, USA

[5]Department of Integrative Biomedical Sciences, Computational Biology Division, University of Cape Town, Cape Town, South Africa

*Corresponding author contact: Egon Ozer (e-ozer@northwestern.edu) and Arghavan Alisoltani (arghavan.alisoltanidehkordi@northwestern.edu)

Abstract

The unprecedented growth of publicly available SARS-CoV-2 genome sequence data has increased demand for effective and accessible SARS-CoV-2 data analysis and visualization tools. A majority of the currently available tools either require computational expertise to deploy or limit user input to pre-selected subsets of SARS-CoV-2 genomes. To address these limitations, we developed ViralVar, a publicly available, point-and-click webtool that gives users the freedom to investigate and visualize user-selected subsets of SARS-CoV-2 genomes obtained from the GISAID public database. ViralVar has two primary features that enable: 1) visualization of spatiotemporal dynamics of SARS-CoV-2 lineages, and 2) structural/functional analysis of genomic mutations. As proof-of-principle, ViralVar was used to explore the evolution of the SARS-CoV-2 pandemic in the USA in the pediatric, adult, and elderly population (n > 1.7 million genomes). While the spatiotemporal dynamics of variants did not differ between these age

groups, several USA-specific sublineages arose relative to the rest of the world. Our development and utilization of ViralVar to provide insights on the evolution of SARS-CoV-2 in the USA demonstrates the importance of developing accessible tools to facilitate and accelerate large-scale surveillance of circulating pathogens. The ViralVar webserver is freely available at http://viralvar.org/.

Keywords: evolution; mutation; genomic surveillance; SARS-CoV-2; COVID-19; ViralVar; webtool

INTRODUCTION

Since the onset of the coronavirus disease 2019 (COVID-19) pandemic, the continued mutation and diversification of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has resulted in the repeated emergence of new 'variants of concern' (VOCs) with increased infectivity, transmissibility and/or immune evasion properties [1-5]. Each VOC has been defined by a distinct set of protein mutations (missense or non-synonymous substitutions, in-frame insertions, and deletions) that confer unique functional properties [1, 6-12]. For example, the Alpha (B.1.1.7*) VOC was defined by a set of nine Spike mutations (N501Y, A570D, D614G, P681H, T716I, S982A, D1118H, 69-70Δ, 144Δ) that increased infectivity [13], transmissibility [14], and resistance to monoclonal antibody therapeutics [15]. Especially within the Spike open reading frame, a greater proportion of missense compared to synonymous mutations is indicative of strong positive selection for Spike proteins with altered structure and function [16, 17]. Continued SARS-CoV-2 genomic surveillance is essential to identify new emergent variants with novel phenotypic properties that may alter best practices in public health and clinical care.

The remarkable global scientific response to the COVID-19 pandemic has led to the generation of vast amounts of publicly available SARS-CoV-2 whole-genome sequence data. Worldwide, most genome sequences are deposited in the GISAID initiative public database (gisaid.org) [18], and more than 13 million viral sequences from around the world have been deposited as of 12th September, 2022. This massive and ongoing SARS-CoV-2 sequencing effort has provided a unique opportunity to study the virus' evolution in exquisite detail. However, at the same time, the volume and diversity of available sequences exacerbates the complexity of the data analysis and calls for effective tools to allow researchers with little or no computational expertise to perform detailed analyses of relevant genomic data.

In part to address this problem, several web-based tools have been developed to facilitate the study of SARS-CoV-2 spatiotemporal dynamics, mutational frequency, and/or three-dimensional (3D) protein structure [19-22]. Though useful for gaining broad insights, these applications are often limited to analysis pre-determined datasets with minimal user control such as, COVIDCG [23], outbreak info [24], covariants [25], 2019nCoVR [26], CoV-GLUE [27] and COG-UK [28]. However even tools that allow processing of user defined data often accept limited number of sequencing data such as covdb (limit= 100) [29], coronApp [30] (limit ~100 MB or ~3500) and VirusViz (limit=50) [31]. In addition to the lack of options for large-scale data analysis, these tools have limited analytical features for multilevel analysis and visualization of SARS-CoV-2 lineages and their mutations (e.g., spatiotemporal visualization of lineages, linear or 3D visualization of mutations in the context of proteins and genome).

Other tools and databases have been developed to study SARS-CoV-2 protein structures. One of these applications is SARS-CoV-2 3D [32] which provides tools for 3D structure predictions and energy calculations to evaluate targets and design new potential therapeutics. CoV3D [33] is a repository for 3D protein structures of SARS-CoV-2 and host antibodies. Neither tool provides information on mutational changes in the context of the 3D structures. Other webservers such as GISAID [18], covariants [25], and COG-UK [28] provide limited 3D structural visualizations for only fixed sets of mutations (mostly clade defining) and only for the Spike protein. To the best of our knowledge there are currently two webservers that enable the visualization of mutations in the context of 3D protein structure for all SARS-CoV-2 proteins: Coronavirus3D [19] and COVID-3D [21]. However, both servers have a fixed list of SARS-CoV-2 genomes/mutations in their databases and lack the option to visualize mutations based on user-provided genomes. Although COVID-3D [21] provides an option for the user to input variants, this is limited to only one protein at a time and thus requires multiple file uploads to visualize genomic mutations in the context of more than one protein. Moreover, both servers lack the capacity to examine mutational patterns at selected time points in the pandemic, in selected geographical regions, and/or among specific lineages.

Here, we present a new, web-based software application ‘ViralVar’ that incorporates user-selected genome data to visualize and study lineages over time by depicting the distribution of mutations at both the nucleotide and protein levels as well as providing the context of variants in the 3D structure of SARS-CoV-2 proteins. Protein visualizations provide detailed information on functional protein domains and predicted B-cell epitopes. Additionally, ViralVar provides a currently unique feature among similar applications that allows for the binomial test of protein

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

mutations to identify potential over- and under mutated proteins, k-means clustering of genomes based on protein mutations to expedite large-scale surveillance of new mutations, and visualization of changes in the mutational patterns of the virus over selected date ranges, within defined geographical regions, and/or within or among lineages. A practical demonstration of the application of ViralVar is given here by examining the relative dynamics of SARS-CoV-2 evolution in the USA using a total of 1,739,797 sequenced genomes collected in the USA between Jan 2020 and May 2022.

METHODS

General Software Workflow

The ViralVar webtool is implemented in the R programming language using Shiny, an open-source R package for developing interactive web applications. Shiny implements layout features available in Bootstrap, an HTML 4.01/ shiny-css 1.7.1/shiny-javascript 1.7.1 framework. To add more advanced content to ViralVar, the user interface was customized with HTML and Shiny's HTML tag attributes, as well as custom Cascading Style Sheets (CSS) and other R packages listed in the context of the relevant sections below. Briefly, SARS-CoV-2 genomic data retrieved from GISAID [18] is used as input for ViralVar. The webapp is divided into two modules (**Figure 1**). In the first module, 'Lineage Dynamics', data are processed to depict spatiotemporal dynamics of SARS-CoV-2 lineages and clades in the form of stacked bars, area plots, and pie charts. The second module, 'Mutational Analysis' visualizes mutation distributions along the SARS-CoV-2 genome and proteins (linear and 3D) and also generates statistical analyses to identify over- and under-mutated proteins. Users can interact with the server to explore and compare the temporal dynamics of lineages and mutations between different sets of genomes and/or VOCs. Each module provides various control options, allowing users to customize analyses and view and export figures according to their requirements. Output files of ViralVar are either high-resolution figures (PDF, PNG) or tables (tsv).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

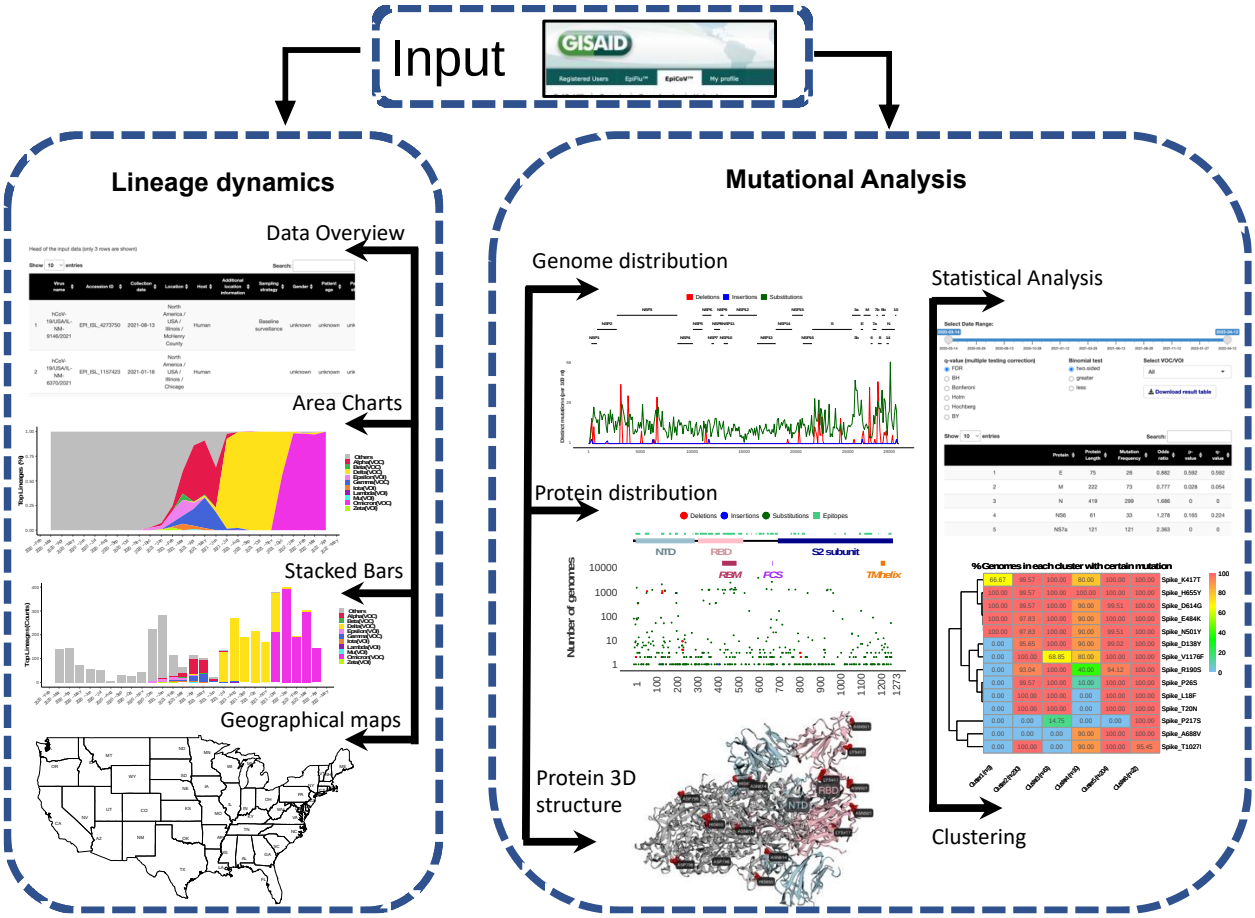

Figure 1 | General workflow of ViralVar and its two main modules. Input data reflecting SARS-CoV-2 sequences of interest can be downloaded directly from the GISAID public repository. In the 'Lineage Dynamics' module, the spatiotemporal dynamics of SARS-CoV-2 lineages and clades are represented in the form of stacked bars, area plots, and pie charts over user selected timeframes and geographical areas. In the 'Mutational Analysis' module, mutations are depicted in the context of the SARS-CoV-2 genome and relevant proteins (both primary sequence and 3D structural representations). This module also provides options to perform statistical analyses to identify over- and under mutated proteins over user selected time periods and to perform genome clustering within user selected subsets. More details are available in the ViralVar User Manual.

Data Input

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

The GISAID Initiative is one of the largest global resources for sharing SARS-CoV-2 genome sequences and associated clinical and demographic information [18]. GISAID data are accessible to users through free registration via the GISAID website (https://gisaid.org/). The database provides genome consensus sequences, reference-based multiple sequence genome alignments, and lists of mutations for each genome with associated lineage or clade designations in tabular format. Data to be downloaded from the database can be readily filtered to focus on dates of collection, specific geographical regions, or selected lineages or clades. ViralVar accepts input data from GISAID in tabular format that includes three sets of information for each genome: 1) PANGO lineage (users can opt to manually add Nextclade designations), 2) sample collection date and 3) a list of protein mutations (denoted as 'AA Substitutions' in GISAID data; required for 'Mutational Analysis'). Detailed guidance on retrieving GISAID data in the proper format for input into ViralVar is provided on the 'Home' tab of ViralVar. A limited set of 3,892 SARS-CoV-2 sequences collected through the Northwestern Medicine Healthcare (NMH) system in Chicago, IL between February 2020 and May 2022 are included in the ViralVar webtool for example purposes and can be viewed by checking the 'Visualize Example Data' checkbox in each module. GISAID IDs are provided in Table S1.

Lineage Dynamics

The "Lineage Dynamics" module of ViralVar serves to provide tools for visualizing changing trends in SARS-CoV-2 lineages/clades over time using temporal abundances and geographical distributions. ViralVar uses the R package ggplot2 [34] to generate visualizations reflecting the trend of changes in absolute and relative abundances of SARS-CoV-2 lineages over time in the input data set. After data input, the data is displayed in tabular format in the 'Data Overview' tab. Note that for this module, only collection date and PANGO lineage information is required. The 'Area Charts' and 'Bar Charts' tabs illustrate the dynamics of lineage distributions over user-specified date ranges. The 'Geographical map' tab shows lineage distributions overlaid as pie charts on user-selected geographical maps for the world, USA, or individual USA states and territories again over a user-specified date range. Geographical maps are drawn using the R packages maps and scatterpie. The phylogenetic nomenclature option allows users to customize output data to use PANGO lineage, Nextclade clade, or World Health Organization-defined VOC nomenclature. Tables and customizable figures are downloadable in Portable Document Format (PDF).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

Mutational Analysis

The "Mutational Analysis" module of ViralVar provides users with a suite of tools to visualize the genomic and structural context of SARS-CoV-2 mutations. The R package ggplot2 [34] is used to generate and annotate density plots. After data input, the data is displayed in tabular format in the 'Data Overview' tab. Note that for this module, collection date, PANGO lineage, and amino acid (AA) substitution information are required. The 'Genome Distribution' tab depicts mutation density among uploaded sequences across the SARS-CoV-2 genome. Briefly, the number of distinct mutation events at each genomic position or protein residue is determined relative to a reference sequence (NCBI: NC_045512.2) [35] and reported over a sliding 100 nucleotide window. Position counts are calculated separately for insertions, deletions, and substitutions. This method does not consider virus counts in its calculation (*i.e.*, the number of uploaded genomes with a particular mutation) such that each mutation event is counted only once. This avoids potential biases in reporting mutational frequency due to unequal amplification or sequencing across the genome as well as bias sampling [16]. In the 'Protein Distribution' tab, frequencies of genomes (virus counts) with mutations at specific protein residues are visualized using the R package ggplot2 [34] and plotly R package (interactive visualization). Separate plots can be generated for all SARS-CoV-2 proteins, both structural and non-structural. Protein domain boundaries are indicated as described in the literature [16, 17]. The IEDB server (Bepipred Linear Epitope Prediction 2.0 at http://www.iedb.org/) [36] was used to predict B-cell epitopes, which are indicated above the protein schematic. In the '3D Protein Structure' tab, the R library package r3dmol is used to visualize mutations in the context of 3D protein structures. 3D coordinates were obtained from the Protein Data Bank (PDB) with PDB accession numbers provided for each structure [37]. For proteins with no available 3D structure, models as predicted by AlphaFold were used when available [38]. Alternatively, the positions of transmembrane helices for proteins with no available 3D structures were identified with the TMHMM 2.0 algorithm [39]. Lists of top mutations along with their frequencies for each protein can be downloaded in the form of tab delimited tables. 3D protein illustrations can be downloaded as Portable Network Graphics (PNG) files. Each of the above tabs includes a date slider to allow users to restrict data to a specific date range and a 'Select VOC/VOI' option to limit output to a specified VOC or VOI.

The above mutational analysis tabs are further complemented by two tabs for statistical analysis and k-means clustering. In the 'Statistical Analysis' tab, ViralVar utilizes the binomial test to identify individual proteins within the uploaded dataset that have significantly different mutation frequencies. The method has been previously applied to identify significantly under- and over-

mutated SARS-CoV-2 proteins [16, 17]. Briefly, the arguments for the binomial test are the observed number of distinct protein mutations in a certain protein (the “number of successes”), the total number of distinct protein mutations in all SARS-CoV-2 proteins (the “number of trials”), and the length of a given protein divided by the length of all SARS-CoV-2 proteins (the “expected probability of success”). An example of binomial calculations is provided below. For more detail please refer to [16].

To simplify the calculations, in this method we hypothesize that each protein mutation is an independent event, and that all SARS-CoV-2 proteins and all residues have the same probability of being mutated. Therefore, this method applies the binomial test to assess the null hypothesis: that protein mutations are distributed randomly across all SARS-CoV-2 proteins.

$$P(MP, MT) = \binom{MT}{MP} P(p)^{MP} (1 - (P(p))^{MT-MP}$$

MT= the total number of protein mutations observed for all proteins (for example, 325 mutations in user input data)

MP= the number of protein mutations in the target protein (for example, 66 mutations in Spike in user input data)

P(p)= Length protein/Length proteome (e.g., length Spike/ total length = 1273/9930=~0.13)

$P(MP, MT) = \binom{325}{66} 0.13^{66} (1 - (0.13)^{325-66}$*=0.00046*

Given that 66 out of the total 325 mutations identified in SARS-CoV-2 proteins are located in Spike, as the length of the Spike protein is of 1273 amino acids and the entire SARS-CoV-2 proteome is 9930 long, based on the null hypothesis we expect only 42 mutations in Spike. However, the binomial test p-value (0.00046) suggests rejection of the null hypothesis and indicates a significantly higher number of mutations in Spike protein compared to the background (entire proteome). ViralVar conducts the above calculation for user input data, therefore MP, MT and *P(p)* will be different for each input dataset. An option to exclude clade signature mutations is provided to avoid bias in the binomial test across highly divergent clades. ViralVar also provides control options to customize binomial test parameters, including the option to adjust p-values for multiple comparisons. As above, the tab includes a date slider to allow users to restrict data to a specific date range and a ‘Select VOC/VOI’ option to limit output to a specified VOC or VOI. A results table of the analysis can be downloaded as a tsv file.

In the ‘Genome Clustering’ tab, ViralVar employs k-means clustering to facilitate rapid investigation of emerging clusters of genomes with specific protein mutation. Since the selection

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

of mutations in SARS-CoV-2 evolution has been shown to be largely impacted by positive selection driven by changes in SARS-CoV-2 protein structure and function [16, 17], targeting protein mutations could cluster genomes relative to phenotype. For instance, a common feature of SARS-CoV-2 genomes with the N501Y spike mutation (e.g., Alpha, Beta and Gamma strains) was enhanced infectivity and transmissibility over previous variants [14].

Clustering of genomes based on pairwise distance-based methods is computationally intensive and might take days to run depending on the computational resources. The runtime for the first step of these approaches (calculation of distance matrices for all pairs of genomes) increases exponentially with the increase in the number of genomes (**Figure S3**). In contrast, K-means clustering of SARS-CoV-2 genomes has been proposed in recent literature as a rapid method to investing emerging variants and tackle computational challenges in large-data analysis [40, 41]. Due to simplicity and being computationally inexpensive, k-means clustering of genomes based on mutations in specific proteins can be quickly and repeatedly run on large-scale genomic datasets (such as ~11.1 M SARS-CoV-2 genomes).

ViralVar uses k-means to group genomes-based on protein mutations. To avoid the effects of spurious mutations (e.g., due to sequencing or assembly errors), the clustering of the genomes is calculated only from protein mutations with a default Minimum Mutation Frequency (MMF) of > 0.005, though this cutoff is user-adjustable. To determine the optimal number of clusters, ViralVar repeats k-means clustering for numbers of clusters (determined based on the number of variables in the input file) and calculates Average Silhouette Width (ASW) index using the R package NbClust [42]. In the calculation of the ASW, ViralVar uses unique genomes (duplicated genomes with identical mutational patterns are removed) to make calculations less computationally expensive. However, the final clustering is applied to all of the genomes in the input data to produce counts of genomes in each cluster. As with previously-described functions, VOC/VOI and date range are selectable. The protein selection option allows for targeting mutations along a protein of interest. Tables and customizable figures in PDF format are downloadable.

Applying ViralVar to assess dynamics of SARS-CoV-2 evolution in the USA

A total of 1,739,797 SARS-CoV-2 high quality complete genome assemblies (GISAID criteria, including N content <5%) from the USA for which patient age, collection date (month/day/year), and geographic location were available were retrieved from GISAID for collection dates between January 1, 2020 and May 15, 2022 (downloaded May 31, 2022). To

study the dynamics of SARS-CoV-2 evolution using ViralVar, sequenced samples were classified into three populations by age: children (0-18 years), adults (18–64 years), and elderly (65+ years) (**Table 1**). The list of GISAID identifiers that compose each group is provided in **Table S1**. Sequence data for each age group was uploaded separately and analyzed using ViralVar. Mutation distributions were also compared between SARS-CoV-2 genomes collected and sequenced for different age groups in the USA.

Table 1 | Details of SARS-CoV-2 data used in this study. Data retrieved from GISAID and each of the three data subsets separately analyzed using ViralVar.

	Sequences	Mean Age	Median Age
Children (<18)	282,106	10.22	10.5
Adults (18-65)	1,287,058	38.92	37.5
Elderly (>65)	170,633	74.42	72.5

RESULTS and DISCUSSION

Spatiotemporal dynamics of SARS-CoV-2 VOCs in the USA

The United States has experienced one of the world's highest COVID-19 burdens during the pandemic, with a total of 86.4M confirmed cases and 1.01M deaths as of May 31, 2022. While some reports are available detailing the evolution of the COVID-19 pandemic in select cities and states [43, 44], there are few comprehensive reports at the national level. To demonstrate the capabilities of ViralVar, we downloaded all high quality whole genome sequence data available in GISAID on specimens collected in the USA between January 1, 2020 and May 15, 2022 (n = 1,739,797 SARS-CoV-2 sequences total). These data were sorted by age (children ,adults, elderly) and uploaded into the ViralVar webtool for analysis.

Temporal dynamics of VOCs across age groups in the USA were visualized using the 'Area Chart' tab in 'Lineage Dynamics' module of ViralVar (**Figure 2A**). Results indicate the dynamics of VOCs were relatively similar for all age groups. SARS-CoV-2 lineage B.1.1.7, designated by the WHO as 'Alpha', was the first named VOC and likely emerged in the United Kingdom (UK) in September 2020. Alpha rapidly displaced other circulating lineages in the USA and became one of the top circulating VOCs in the world in early 2021 [45]. The emergence of Alpha in the USA can be tracked back to November 2020 (**Figure 2B**), coincident with a spike of new cases and deaths between November 2020 and March 2021 (**Figure S1**). Using the date range feature to focus on dynamics during these months, Alpha emerges as the dominant variant

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

at the tail end of the surge in cases, suggesting that it was not responsible for the rise in cases, but rather took over after contraction of cases of the previous variant (**Figure 2B**).

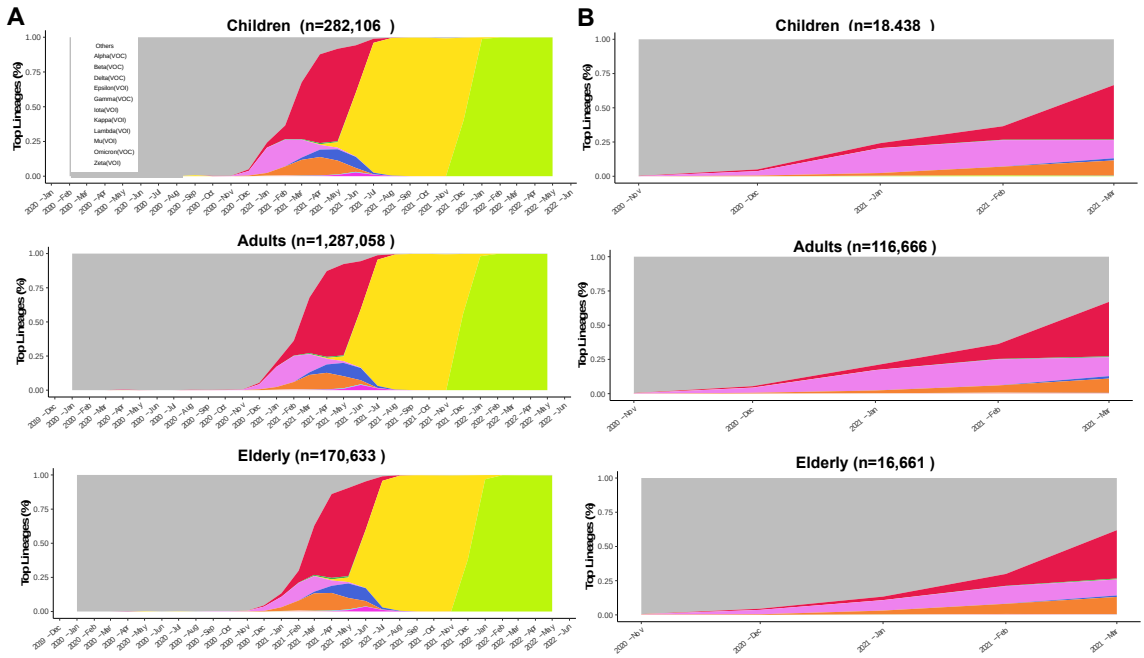

Figure 2 | Area plots reflecting the relative abundance of variants of concern and variants of interest collected in the USA over time. A) Frequency of indicated VOCs and VOIs over time in specimens collected between January 2020 and May 2022 in the USA (n = 1,739,797 sequences, from GISAID as of May 31, 2022). **B**) Frequency of indicated VOCs and VOIs over time in specimens collected between November 2020 and March 2021. Specimens were divided into three age groups: children (up to 18 years), adults (18–64 years), and the elderly (65 years or more). The number of sequences per age group is indicated above each plot. Each subset of genomes was processed separately using the ViralVar 'Lineage Dynamics' module.

Utilizing the 'Geographical Map' feature in 'Lineage Dynamics' module, the distribution of VOCs collected between January 2020 and May 2022 was visualized for each age group by state (**Figure 3A**). Lineage distributions were similar across states between all age groups, with the Omicron and Delta VOCs making up a majority of cases, followed by Alpha (**Figure 3A, Figure S2**). In narrower timeframes, however, distinct spatiotemporal trends become more obvious. Using the date control feature, we adjusted this analysis to examine cases between November 2020 and March 2021 (**Figure 3B**). While Alpha lineages made up a majority of cases in most

states over this time period, region-specific trends emerged. For example, the Epsilon VOI was responsible for a substatial number of cases in southwestern states, while the Iota VOI was more prevalent in the northeast. Illinois specifically reported a substantial number of cases of the Gamma VOC that were not reflected in neighboring states. These region-specific trends were consistent across age groups (**Figure 3B**).

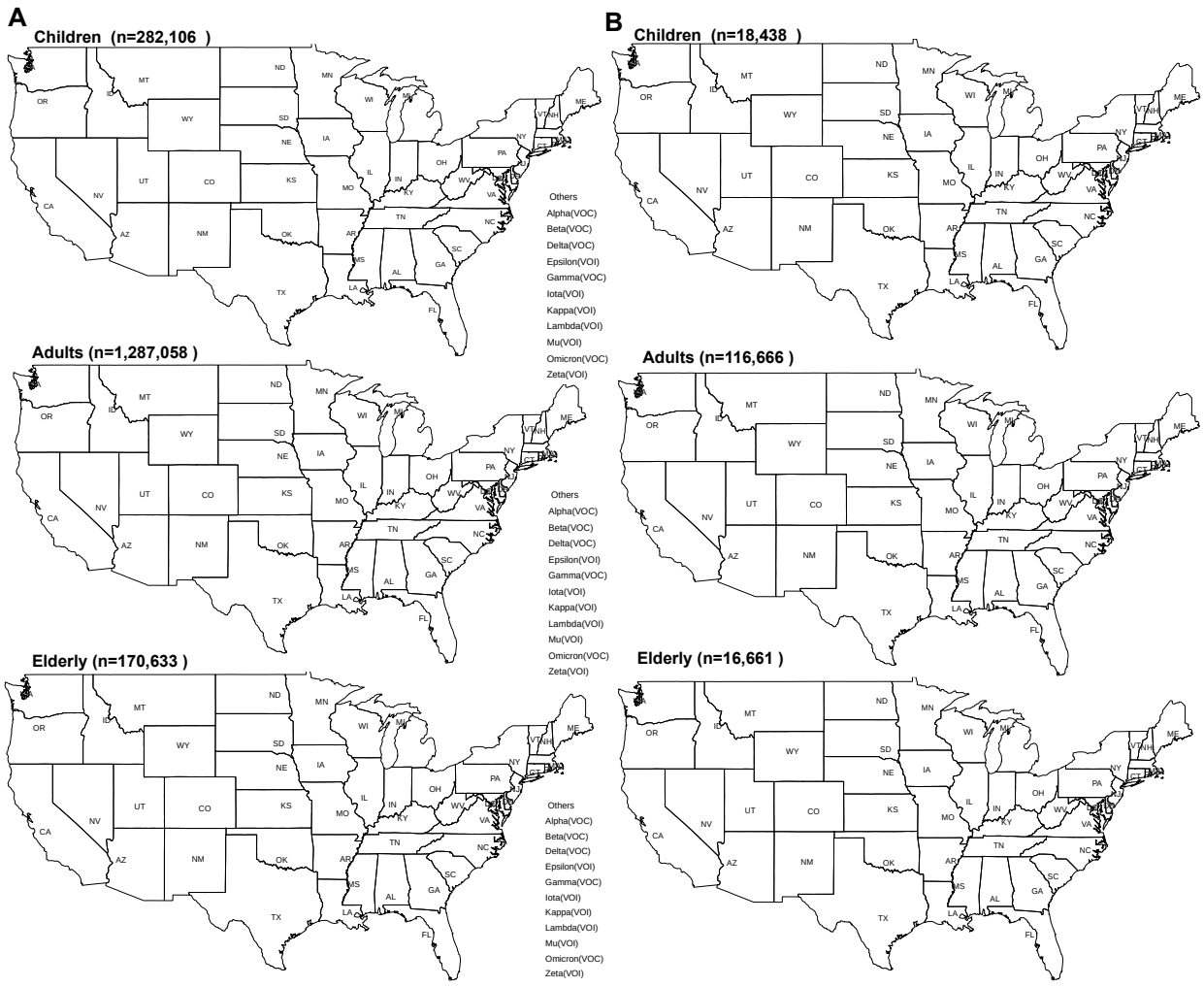

Figure 3 | Distribution of SARS-CoV-2 variants of concern and interest by US state. Pie charts represent the proportion of SARS-CoV-2 VOIs and VOCs in each US state as reported to GISAID (as of May 31, 2022) **A)** between January 2020 and May 2022, and **B**) between November 2020 and March 2021. Specimens were divided into three age groups: children (up to 18 years), adults (18–64 years), and the elderly (65 years or more). The number of sequences per age group is indicated above each plot. The size of pie charts represents the relative

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

frequency of sequenced data in each state. Each subset of genomes was visualized separately using the ViralVar 'Geographical Map' feature.

Mutational analysis of Alpha variant sublineages in the USA

We subset the USA data explained earlier (n = 1,739,797 SARS-CoV-2 sequences) to only include genomes assigned to the Alpha lineage (n=140,100). Additionally, genomes assigned to the Alpha lineage collected from entire world (n=906,114 excluding the USA) were retrieved from GISAID as of May 31, 2022. Using the ViralVar 'Mutational Analysis' module, the mutation profile for the Alpha VOC in the USA was compared to specimens from other countries. All ages were grouped together for this analysis due to the relatively small sample size of the under 18 and over 65 populations compared to adults in this data set. Using the 'Protein Distribution' tab, we visualized the mutational frequency in Alpha VOC sequences at sites across Spike and NSP12 in both the USA and in the rest of the world (**Figure 4**). While the defining mutations of the Alpha VOC are universally present, a distinct subset of mutations were more prevalent in the USA, specifically Spike mutation K1191N and NSP12 mutation P227L. To further investigate these mutations, all genomes containing Spike K1191N (n = 51,713) and NSP12 P227L (n = 190,869) mutations were retrieved from GISAID and uploaded into ViralVar. A majority of genomes with the Spike K1191N mutation were in Alpha variant genomes (80.4%, 41,558 of 51,713 genomes), of which the vast majority came from the USA (93.5%, 38,837 of 41,558 genomes) (**Figure 5A**, top). Similarly, 169,314 genomes with the NSP12 P227L mutation were classified as Alpha variants (88.7%), of which 104,435 genomes were collected in USA (61.6%) (**Figure 5A**, bottom).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

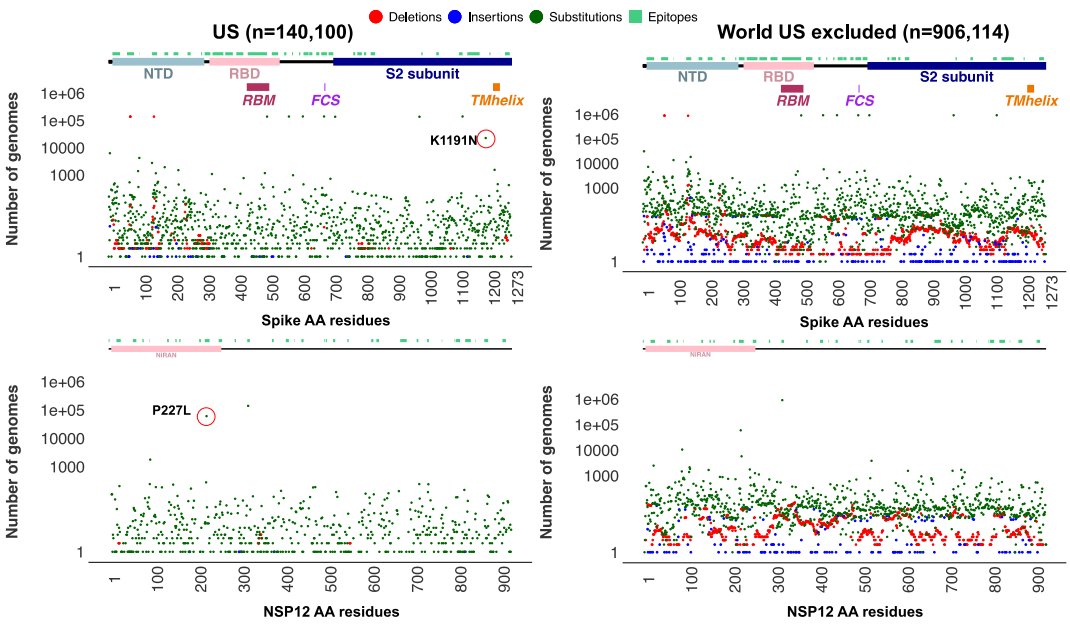

Figure 4 | Absolute frequency of mutations in SARS-CoV-2 Spike and NSP12 among Alpha VOCs. SARS-CoV-2 genome data for all sequences assigned to an Alpha variant lineage (B.1.1.7 and Q.*) from the USA (n = 140,100) and rest of the world (n = 906,114, USA cases excluded) were retrieved from GISAID as of May 31, 2022. Plots represent the absolute frequency of mutations at each amino acid position across Spike (top) and NSP12 (bottom) in sequences form the USA (left) and rest of the world (right). Deletions (red), insertions (blue), and substitutions (green) are plotted in different colors at each position. Boundaries for protein domains of Spike and NSP12 proteins are obtained from [16, 17]. Predicted B cell epitopes are highlighted above in teal as predicted by [68]. Each subset of genomes was visualized separately using the ViralVar 'Protein Distribution' feature.

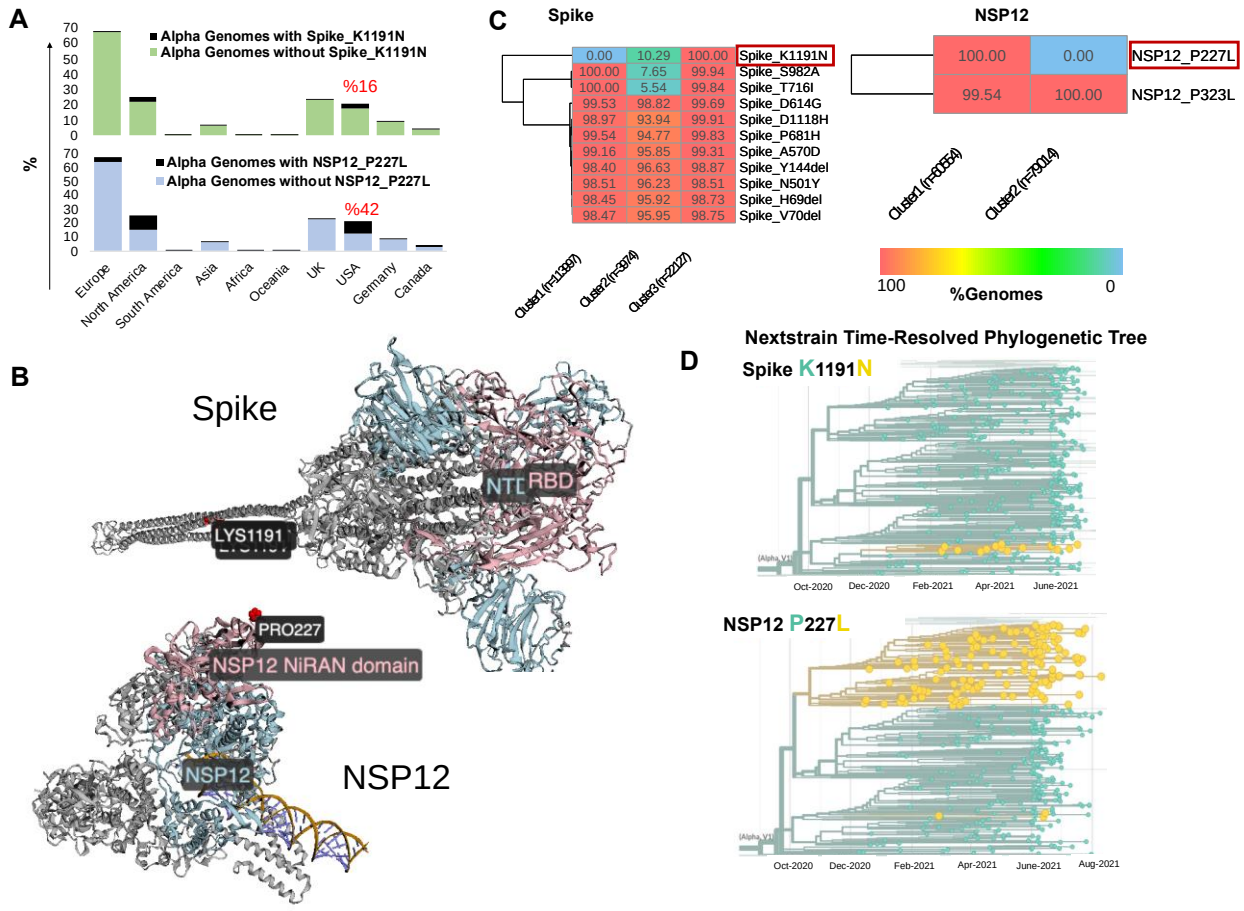

Figure 5 | Analysis of Alpha VOC mutations predominantly found in USA specimens. A) Relative frequency of Alpha VOC SARS-CoV-2 genomes harboring Spike K1191N (top) or NSP12 P227L (bottom) mutations. Calculations are based on GISAID data as of May 31, 2022. **B)** Spike K1191N (top) or NSP12 P227L (bottom) mutations highlighted on available protein structures using the ViralVar '3D Protein Structure' feature. The Spike receptor-binding domain (RBD) and N-terminal domain (NTD) are colored in light blue and pink, respectively (D-I-TASSER model). NSP12 is colored in light blue with the NiRAN domain highlighted in pink (PDB: 6XEZ). **C)** Euclidean distance-based k-means clustering of Alpha VOC SARS-CoV-2 genomes based on Spike and NSP12 mutations was performed using the 'Genome Clustering' feature. Heatmaps represent the percent of genomes with a specific mutation within each cluster. Only protein mutations present in more than two thirds (70%) of genomes are shown here. **D**) Time-resolved phylogenetic tree built by Nextstrain (https://nextstrain.org/ncov/gisaid/north-america/) using a North America-focused subsampling between Dec 2020 and Aug 2021 (n= 399 sequences). Yellow branches and tips highlight genomes containing the Spike K1191N (top) and NSP12 P227L (bottom) mutations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

Spike K1191N and NSP12 P227L appear to be recurrent mutations that have emerged in several other VOCs (*i.e.*, Delta, Omicron and Gamma); however, there is a lack of evidence regarding their role in virus infectivity, transmissibility, and/or clinical outcomes. To gain insight into their possible functional roles, we examined the protein context of each mutation using the '3D Protein Structure' feature in ViralVar (**Figure 5B**). The NSP12 P227L mutation is located in the Nidovirus RdRp associated nucleotidyl transferase (NiRAN) domain. While it is surface exposed, it is far from the RNA binding or enzymatic active site. That being said, a nearby mutation in the NiRAN domain, N198S, has been recently reported as a potential antiviral resistance mutation to the NSP12-targeting drug, remdesivir [46]. Given the high level of conservation among coronavirus RNA-dependent RNA polymerases (RdRps) [47] and the recurring, but infrequent, prevalence of this mutation, it may also be that the P227L mutation confers some selective benefit, but at a fitness cost to the virus [46]. Spike mutation K1191N is located in the S2 subunit in the heptad repeat 2 (HR2) subdomain of the Spike protein, which is involved in host cell membrane fusion and viral entry (**Figure 5B**) [48]. Other Spike protein mutations in the HR2 subdomain such as V1176F have been shown to augment the stability of Spike and have been associated with increased disease severity and mortality [49-51]. More studies are required to determine the functional consequences of both Spike K1191N and NSP12 P227L.

ViralVar k-means clustering feature identifies subclusters of the Alpha variant in the USA

To better understand the genomic context of these mutations, we used the 'Clustering Analysis' feature in ViralVar to identify co-occurring groups of mutations. K-means clustering based on Euclidean distance was applied to all Alpha VOC sequences collected in the USA using a minimum mutation frequency cutoff of 0.005 and with a focus on the Spike and NSP12 proteins. Clustering of Alpha genomes based on Spike mutations resulted in three distinct clusters (**Figure 5C**), two of which were defined by the presence (cluster 3) or absence (cluster 1) of K1191N. A third cluster had a minor presence of K1191N, but concurrently lacked the S982A and/or T716I mutations (cluster 2). Clustering of Alpha genomes based on NSP12 mutations identified two distinct clusters distinguished solely by the P227L mutation (**Figure 5C**). To determine if these clusters are also identified by phylogenetic analysis, we also examined these mutations using the Nextstrain webserver (**Figure 5D**). The time-resolved phylogenetic trees from Nextstrain suggest that the Spike K1191N mutation is monophyletic, while the P227L mutation arose in at least two

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

distinct branches (**Figure 5D**). The k-means clustering is largely in accordance with the phylogenetic analysis, but suggests that additional mutational information, including synonymous mutations and those that occur outside of the open reading frame of interest, capture additional information not accounted for in this approach.

One of the limitations of phylogenetic tree-based analysis, clustering, and visualization of SARS-CoV-2 genomes and investigating protein mutations is the computational cost that multiplies with the number of available genomes. A majority of studies using phylogenetic trees to study SARS-CoV-2 variants of concern (VOCs) therefore must rely on subsampling approaches [52, 53]. K-means-based clustering of SARS-CoV-2 genomes based on Euclidean distance is one way to overcome this challenge since the method calculates distance of each datapoint to centroid instead using pairwise distances, decreasing the computational cost of analyzing additional sequences (**Figure S3**). Furthermore, the k-means clustering of genomes based on protein mutations can be leveraged to group genomes in a way directly related to phenotype [54, 55]. The congruence between the approach taken by ViralVar (**Figure 5C**) and the phylogenetic analysis results (**Figure 5D**) support the potential use of k-means clustering for rapid analysis of large genomic datasets to facilitate tracking emerging protein mutations using a generic clustering method. This method could also be readily adapted and applied to other viruses. That being said, this approach is not suitable for making specific evolutionary inferences and so can be considered complementary to traditional phylogenetic-tree-based methods and useful for initial analyses and hypothesis generation.

Significant non-random distribution of mutations in SARS-CoV-2 proteins

To explore the different mutational profiles in genomes collected for different age groups in the USA, we used the ‘Genome Distribution’ feature of ViralVar to visualize mutations in all collected specimens from the USA split by age group (**Figure 6**). Overall, analyses of mutation profiles of SARS-CoV-2 genomes were relatively similar for the three age groups in the USA samples (**Figure 6**). Compared to structural and accessory proteins, non-structural proteins seem to undergo a higher mutational constraint (**Figure 6** and **Tables S2**), consistent with the previous reports [16]. Slight variability in mutational patterns between different data subsets could be partly attributed to the differences in the population size and sampling dates between regions and age groups.

One of the most noteworthy differences when comparing results from the first year of the pandemic [16] and results obtained in this study is the increased frequency of protein indel events,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

especially the accumulation of insertions in the Spike NTD. This trend was consistent for samples collected across all age groups, though distinct deletion events appeared more prevalent in elderly populations (for example, in the NSP15 open reading frame, **Figure 6**). The increased frequency of recurrent indels and their non-random distribution is believed to be an adaptive response mechanism to elevated global herd immunity resulting from vaccination, infection, or both [17, 61, 62]. Spike NTD indels could alter neutralizing epitopes in the region and are thought to result in reduced antibody protection against VOCs that harbor these indels [61].

Using the 'Statistical Analysis' feature of ViralVar, we further identified significant accumulations of mutations in mostly structural proteins of SARS-CoV-2 with two exceptions for non-structural proteins (NSP1 and NSP2). Of note, a higher concentration of mutations was observed in NSP1 (average odds ratio = 1.46, q-value = 0 across all age groups), NSP2 (average odds ratio = 1.3, q-value = 0 across all age groups), N (average odds ratio = 1.6, q-value = 0 across all age groups), NS6 (average odds ratio = 1.6, q-value = 0 across all age groups), NS7a (average odds ratio = 3.1, q-value = 0 across all age groups), NS7b (average odds ratio = 1.8, q-value = 0 across all age groups), NS8 (average odds ratio = 3.1, q-value = 0 across all age groups) and Spike (average odds ratio = 1.4, q-value = 0 across all age groups) (**Table S2**). All these proteins are involved in interactions with the host immune system [63-65]. Recurrent NSP1 substitutions and indels have been found to accumulate on the protein surface and near epitope regions [17] and are thought to adversely affect the host's immune response and vaccine efficiency [66, 67]. For instance, NSP1 Δ79-89 induces a lower IFN-I response in the infected Calu-3 cells [67], highlighting the biological importance of mutations in NSP1 and other non-structural proteins. The significantly higher concentration of mutations in specific proteins involved in host immune interactions, the emergence of new types of protein mutations (in-frame indels), and the expansion of mutations to new proteins or protein regions suggest the virus is evolving to combat the host immune system. Taken together, non-random distribution of the mutations in different SARS-CoV-2 proteins suggests proteins undergo different evolutionary pressures driven partly by host immune system.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2022 doi:10.20944/preprints202210.0241.v1

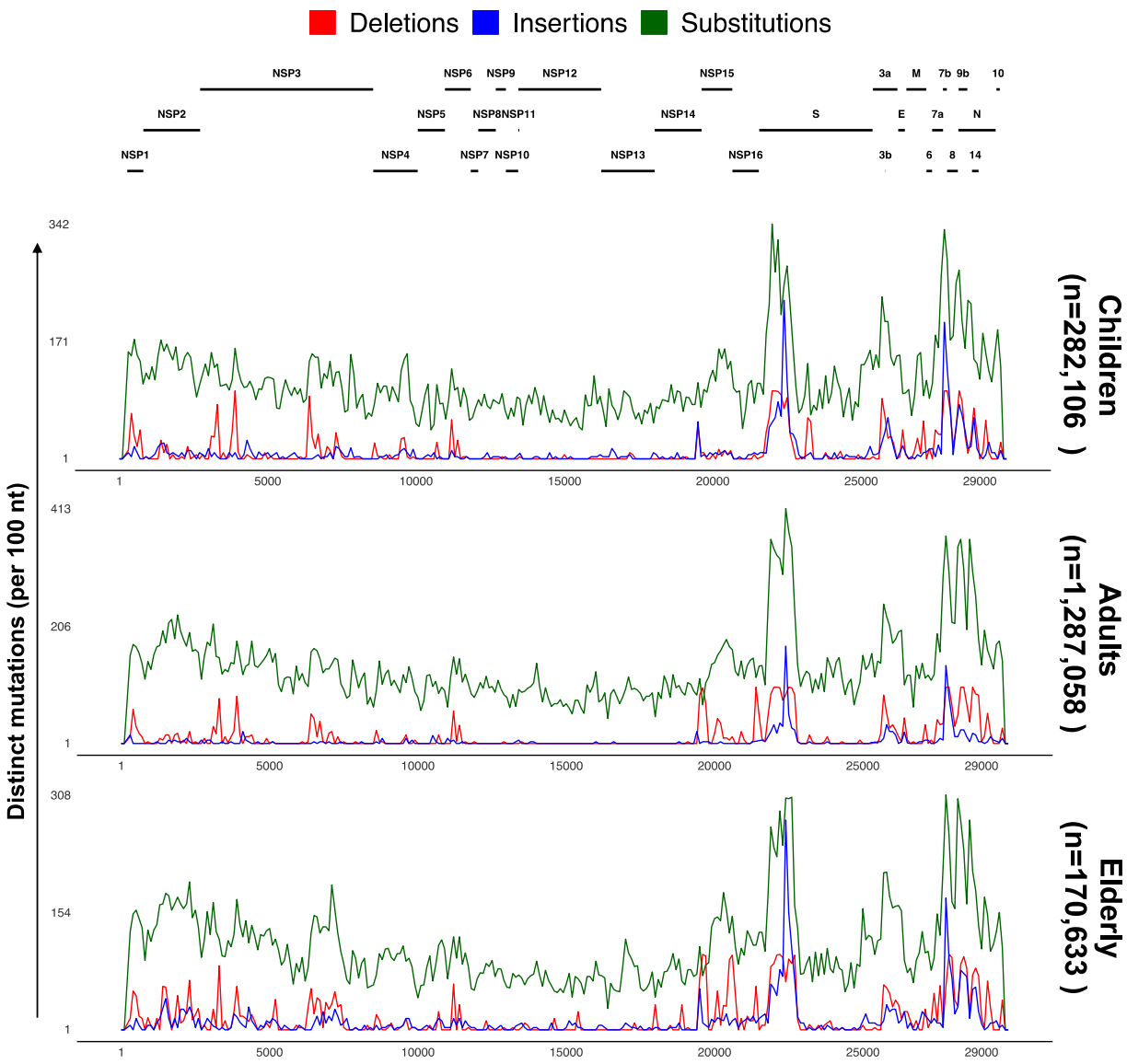

Figure 6 | Genomic distribution of SARS-CoV-2 mutations for three age groups. Each plot depicts the number of distinct protein mutations in a 100 nucleotide sliding window across the SARS-CoV-2 genome in specimens collected between January 2020 and May 2022 in the USA (n = 1,739,797 sequences, from GISAID as of May 31, 2022). Sequences were divided into six groups based on the age of patients [children (up to 18 years), adults (18–64 years), and elderly (65 years or more)]. The total number of sequences used per age group is indicated. Each subset of genomes was processed separately using the ViralVar 'Lineage Dynamics' module.

CONCLUSION

The emergence of new variants of SARS-CoV-2 with higher transmissibility and enhanced immune evasion highlights the need for ongoing SARS-CoV-2 genomic surveillance. This work has been greatly facilitated by public sequence repositories, such as GISAID, which had data available for more than 11.1 M genome sequences as of May 31, 2022. At the same time, this vast amount of genomic data has increased the demand for more flexible and multilevel analysis platforms to help study the virus evolution. To complement and expand upon previously developed analysis tools, we created ViralVar, a webtool for visualizing and researching SARS-CoV-2 lineages and mutational patterns over time. We have shown that ViralVar can be deployed as a point-and-click tool to rapidly investigate the spatiotemporal evolution of large numbers of SARS-CoV-2 genomes. Overall, our findings utilizing ViralVar offer important insights into pathogen evolution dynamics and spread in the USA. This study demonstrates that ViralVar can be successfully used to study the evolution of SARS-CoV-2 and help in improving global COVID-19 mitigation plans as the pandemic continues to evolve.

As part of a larger project for facilitating the study of virus evolution and mutational patterns, development of ViralVar will continue for the study of other viruses. Additional future work includes the addition of multiple data input options (*i.e.*, Consensus sequences or Multiple Sequence Alignments) to facilitate users in analyzing their own data. Continued enrichment of the list of structural and functional properties of SARS-CoV-2 and other viral proteins in ViralVar will also take place on a regular basis. ViralVar databases will be updated at regular intervals based upon information provided for other viruses and updates in public databases for protein structural and functional properties. ViralVar complements current tools for studying the massive number of SARS-CoV-2 genomes and can provide a user-friendly platform for the multilevel study of SARS-CoV-2 evolution.

ACKNOWLEDGEMENTS

We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared via GISAID genetic sequence data on which this research is based, as well as structural biology groups contributing their structures to the PDB. Supported by grants (R21 AI163912 to Dr. Hultquist, U19 AI171110 to Dr. Hultquist, U19 AI135964 to Dr. Ozer) from the National Institutes of Health; a grant (to Dr. Lorenzo-Redondo) from the Northwestern University Havey Institute for Global Health; and a grant (to Drs. Ozer and Hultquist) from the Walder Foundation's Chicago Coronavirus Assessment Network. This research was supported in part through the computational resources and staff contributions

provided for the Quest high performance computing facility at Northwestern University, which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. The funding sources had no role in the study design, data collection, analysis, interpretation, or writing of the report.

AUTHOR CONTRIBUTIONS

Conceptualization, A.A.; Methodology, A.A. L.J., A.I., R.L-R, E.A.O, and AG; Software, A.A.; Data Curation, A.A., L.M.S., T.D.; Writing – Original Draft Preparation, A.A. and E.A.O.; Writing – Review & Editing, A.A., E.A.O., L.M.S., R.L-R., and J.F.H.; Visualization, A.A. L.J., and A.I.; Supervision, E.A.O. and J.F.H.; Project Administration, E.A.O. and J.F.H.; Funding Acquisition, E.A.O. and J.F.H.

CONFLICTS OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY

ViralVar webserver is freely accessible through http://viralvar.org/.

REFERENCES

[1] Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A. J., Engelbrecht, S., Van Zyl, G., Preiser, W., Petruccione, F., Sigal, A., Hardie, D., Marais, G., Hsiao, N. Y., Korsman, S., Davies, M. A., Tyers, L., Mudau, I., York, D., Maslo, C., Goedhals, D., Abrahams, S., Laguda-Akingba, O., Alisoltani-Dehkordi, A., Godzik, A., Wibmer, C. K., Sewell, B. T., Lourenco, J., Alcantara, L. C. J., Kosakovsky Pond, S. L., Weaver, S., Martin, D., Lessells, R. J., Bhiman, J. N., Williamson, C., and de Oliveira, T., 2021, "Detection of a SARS-CoV-2 variant of concern in South Africa," Nature, 592(7854), pp. 438-443.
[2] Karim, S. S. A., and Karim, Q. A., 2021, "Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic," The Lancet, 398(10317), pp. 2126-2128.
[3] Viana, R., Moyo, S., Amoako, D. G., Tegally, H., Scheepers, C., Lessells, R. J., Giandhari, J., Wolter, N., Everatt, J., and Rambaut, A., 2021, "Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa," medRxiv.
[4] Madhi, S. A., Baillie, V., Cutland, C. L., Voysey, M., Koen, A. L., Fairlie, L., Padayachee, S. D., Dheda, K., Barnabas, S. L., Bhorat, Q. E., Briner, C., Kwatra, G., Ahmed, K., Aley, P., Bhikha, S., Bhiman, J. N., Bhorat, A. E., du Plessis, J., Esmail, A., Groenewald, M., Horne, E., Hwa, S. H., Jose,

A., Lambe, T., Laubscher, M., Malahleha, M., Masenya, M., Masilela, M., McKenzie, S., Molapo, K., Moultrie, A., Oelofse, S., Patel, F., Pillay, S., Rhead, S., Rodel, H., Rossouw, L., Taoushanis, C., Tegally, H., Thombrayil, A., van Eck, S., Wibmer, C. K., Durham, N. M., Kelly, E. J., Villafana, T. L., Gilbert, S., Pollard, A. J., de Oliveira, T., Moore, P. L., Sigal, A., Izu, A., and Group, N.-S. G. W.-V. C., 2021, "Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant," N Engl J Med.

[5] Jewell, B. L., 2021, "Monitoring differences between the SARS-CoV-2 B.1.1.7 variant and other lineages," Lancet Public Health, 6(5), pp. e267-e268.

[6] Jassat, W., Mudara, C., Ozougwu, L., Tempia, S., Blumberg, L., Davies, M.-A., Pillay, Y., Carter, T., Morewane, R., and Wolmarans, M., 2021, "Difference in mortality among individuals admitted to hospital with COVID-19 during the first and second waves in South Africa: a cohort study," The Lancet Global Health, 9(9), pp. e1216-e1225.

[7] Edward, P. R., Lorenzo-Redondo, R., Reyna, M. E., Simons, L. M., Hultquist, J. F., Patel, A. B., Ozer, E. A., Muller, W. J., Heald-Sargent, T., and McHugh, M., 2021, "Severity of Illness Caused by Severe Acute Respiratory Syndrome Coronavirus 2 Variants of Concern in Children: A Single-Center Retrospective Cohort Study," Medrxiv.

[8] Duong, D., 2021, "Alpha, Beta, Delta, Gamma: What's important to know about SARS-CoV-2 variants of concern?," Can Med Assoc.

[9] Khan, A., Khan, T., Ali, S., Aftab, S., Wang, Y., Qiankun, W., Khan, M., Suleman, M., Ali, S., and Heng, W., 2021, "SARS-CoV-2 new variants: characteristic features and impact on the efficacy of different vaccines," Biomedicine & Pharmacotherapy, 143, p. 112176.

[10] Wang, Y., Chen, R., Hu, F., Lan, Y., Yang, Z., Zhan, C., Shi, J., Deng, X., Jiang, M., and Zhong, S., 2021, "Transmission, viral kinetics and clinical characteristics of the emergent SARS-CoV-2 Delta VOC in Guangzhou, China," EClinicalMedicine, 40, p. 101129.

[11] Tian, D., Sun, Y., Xu, H., and Ye, Q., 2022, "The emergence and epidemic characteristics of the highly mutated SARS-CoV-2 Omicron variant," Journal of Medical Virology, 94(6), pp. 2376-2383.

[12] Post, L. A., and Lorenzo-Redondo, R., 2022, "Omicron: fewer adverse outcomes come with new dangers," The Lancet, 399(10332), pp. 1280-1281.

[13] Frampton, D., Rampling, T., Cross, A., Bailey, H., Heaney, J., Byott, M., Scott, R., Sconza, R., Price, J., and Margaritis, M., 2021, "Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B. 1.1. 7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study," The Lancet infectious diseases, 21(9), pp. 1246-1256.

[14] Liu, Y., Liu, J., Plante, K. S., Plante, J. A., Xie, X., Zhang, X., Ku, Z., An, Z., Scharton, D., and Schindewolf, C., 2022, "The N501Y spike substitution enhances SARS-CoV-2 infection and transmission," Nature, 602(7896), pp. 294-299.

[15] Lusvarghi, S., Wang, W., Herrup, R., Neerukonda, S. N., Vassell, R., Bentley, L., Eakin, A. E., Erlandson, K. J., and Weiss, C. D., 2022, "Key substitutions in the spike protein of SARS-CoV-2 variants can predict resistance to monoclonal antibodies, but other substitutions can modify the effects," Journal of virology, 96(1), pp. e01110-01121.

[16] Jaroszewski, L., Iyer, M., Alisoltani, A., Sedova, M., and Godzik, A., 2021, "The interplay of SARS-CoV-2 evolution and constraints imposed by the structure and functionality of its proteins," Plos Computational Biology, 17(7).

[17] Alisoltani, A., Jaroszewski, L., Iyer, M., Iranzadeh, A., and Godzik, A., 2022, "Increased frequency of indels in hypervariable regions of SARS-CoV-2 proteins–a possible signature of adaptive selection," Frontiers in Genetics, p. 1019.
[18] Shu, Y., and McCauley, J., 2017, "GISAID: Global initiative on sharing all influenza data - from vision to reality," Euro Surveill, 22(13).
[19] Sedova, M., Jaroszewski, L., Alisoltani, A., and Godzik, A., 2020, "Coronavirus3D: 3D structural visualization of COVID-19 genomic divergence," Bioinformatics, 36(15), pp. 4360-4362.
[20] Mercatelli, D., Holding, A. N., and Giorgi, F. M., 2021, "Web tools to fight pandemics: the COVID-19 experience," Briefings in bioinformatics, 22(2), pp. 690-700.
[21] Portelli, S., Olshansky, M., Rodrigues, C. H., D'Souza, E. N., Myung, Y., Silk, M., Alavi, A., Pires, D. E., and Ascher, D. B., 2020, "Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource," Nature genetics, 52(10), pp. 999-1001.
[22] Mei, L.-C., Jin, Y., Wang, Z., Hao, G.-F., and Yang, G.-F., 2021, "Web resources facilitate drug discovery in treatment of COVID-19," Drug discovery today, 26(10), pp. 2358-2366.
[23] Chen, A. T., Altschuler, K., Zhan, S. H., Chan, Y. A., and Deverman, B. E., 2021, "COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest," Elife, 10, p. e63409.
[24] Gangavarapu, K., Latif, A. A., Mullen, J. L., Alkuzweny, M., Hufbauer, E., Tsueng, G., Haag, E., Zeller, M., Aceves, C. M., and Zaiets, K., 2022, "Outbreak. info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations," Research square.
[25] Hodcroft, E., 2021, "CoVariants: SARS-CoV-2 mutations and variants of interest. 2021 https://covariants. org," Accessed.
[26] Lu, G., and Moriyama, E. N., 2021, "2019nCoVR—A comprehensive genomic resource for SARS-CoV-2 variant surveillance," The Innovation, 2(4).
[27] Singer, J., Gifford, R., Cotten, M., and Robertson, D., 2020, "CoV-GLUE: a web application for tracking SARS-CoV-2 genomic variation."
[28] Wright, D. W., Harvey, W. T., Hughes, J., Cox, M., Peacock, T. P., Colquhoun, R., Jackson, B., Orton, R., Nielsen, M., and Hsu, N. S., 2022, "Tracking SARS-CoV-2 mutations and variants through the COG-UK-Mutation Explorer," Virus Evolution, 8(1), p. veac023.
[29] Tzou, P., Tao, K., Sahoo, M. K., Pond, S. L., Pinsky, B. A., and Shafer, R. W., "Sierra SARS-CoV-2 Sequence and Antiviral Resistance Analysis Program," Available at SSRN 4160017.
[30] Mercatelli, D., Triboli, L., Fornasari, E., Ray, F., and Giorgi, F. M., 2021, "Coronapp: a web application to annotate and monitor SARS-CoV-2 mutations," Journal of medical virology, 93(5), pp. 3238-3245.
[31] Bernasconi, A., Gulino, A., Alfonsi, T., Canakoglu, A., Pinoli, P., Sandionigi, A., and Ceri, S., 2021, "VirusViz: comparative analysis and effective visualization of viral nucleotide and amino acid variants," Nucleic Acids Research, 49(15), pp. e90-e90.
[32] Alsulami, A. F., Thomas, S. E., Jamasb, A. R., Beaudoin, C. A., Moghul, I., Bannerman, B., Copoiu, L., Vedithi, S. C., Torres, P., and Blundell, T. L., 2021, "SARS-CoV-2 3D database: understanding the coronavirus proteome and evaluating possible drug targets," Briefings in bioinformatics, 22(2), pp. 769-780.

[33] Gowthaman, R., Guest, J. D., Yin, R., Adolf-Bryfogle, J., Schief, W. R., and Pierce, B. G., 2021, "CoV3D: a database of high resolution coronavirus protein structures," Nucleic acids research, 49(D1), pp. D282-D287.
[34] Wickham, H., 2011, "ggplot2," Wiley Interdisciplinary Reviews: Computational Statistics, 3(2), pp. 180-185.
[35] Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., and Pei, Y.-Y., 2020, "A new coronavirus associated with human respiratory disease in China," Nature, 579(7798), pp. 265-269.
[36] Jespersen, M. C., Peters, B., Nielsen, M., and Marcatili, P., 2017, "BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes," Nucleic Acids Res, 45(W1), pp. W24-W29.
[37] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E., 2000, "The Protein Data Bank," Nucleic Acids Res, 28(1), pp. 235-242.
[38] DeepMind, "Computational predictions of protein structures associated with COVID-19."
[39] Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L., 2001, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," J Mol Biol, 305(3), pp. 567-580.
[40] Hozumi, Y., Wang, R., Yin, C., and Wei, G. W., 2021, "UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets," Comput Biol Med, 131, p. 104264.
[41] Wang, R., Chen, J., Gao, K., Hozumi, Y., Yin, C., and Wei, G.-W., 2021, "Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants," Communications biology, 4(1), pp. 1-14.
[42] Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A., 2014, "NbClust: an R package for determining the relevant number of clusters in a data set," Journal of statistical software, 61(1), pp. 1-36.
[43] Deng, X., Gu, W., Federman, S., Du Plessis, L., Pybus, O. G., Faria, N. R., Wang, C., Yu, G., Bushnell, B., and Pan, C.-Y., 2020, "Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California," Science, 369(6503), pp. 582-587.
[44] Russell, A., O'Connor, C., Lasek-Nesselquist, E., Plitnick, J., Kelly, J. P., Lamson, D. M., and George, K. S., 2022, "Spatiotemporal Analyses of 2 Co-Circulating SARS-CoV-2 Variants, New York State, USA," Emerging infectious diseases, 28(3), p. 650.
[45] Alpert, T., Brito, A. F., Lasek-Nesselquist, E., Rothman, J., Valesano, A. L., MacKay, M. J., Petrone, M. E., Breban, M. I., Watkins, A. E., and Vogels, C. B., 2021, "Early introductions and transmission of SARS-CoV-2 variant B. 1.1. 7 in the United States," Cell, 184(10), pp. 2595-2604. e2513.
[46] Stevens, L. J., Pruijssers, A. J., Lee, H. W., Gordon, C. J., Tchesnokov, E. P., Gribble, J., George, A. S., Hughes, T. M., Lu, X., and Li, J., 2022, "Mutations in the SARS-CoV-2 RNA dependent RNA polymerase confer resistance to remdesivir by distinct mechanisms," Science translational medicine, p. eabo0718.
[47] Posthuma, C. C., Te Velthuis, A. J., and Snijder, E. J., 2017, "Nidovirus RNA polymerases: complex enzymes handling exceptional RNA genomes," Virus research, 234, pp. 58-73.
[48] Huang, Y., Yang, C., Xu, X.-f., Xu, W., and Liu, S.-w., 2020, "Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19," Acta Pharmacologica Sinica, 41(9), pp. 1141-1149.

[49] Nagy, Á., Pongor, S., and Győrffy, B., 2021, "Different mutations in SARS-CoV-2 associate with severe and mild outcome," International journal of antimicrobial agents, 57(2), p. 106272.
[50] Farkas, C., Mella, A., Turgeon, M., and Haigh, J. J., 2021, "A novel SARS-CoV-2 viral sequence bioinformatic pipeline has found genetic evidence that the viral 3' untranslated region (UTR) is evolving and generating increased viral diversity," Frontiers in microbiology, 12, p. 665041.
[51] Yang, K., Wang, C., White, K. I., Pfuetzner, R. A., Esquivies, L., and Brunger, A. T., 2022, "Structural conservation among variants of the SARS-CoV-2 spike postfusion bundle," Proceedings of the National Academy of Sciences, 119(16), p. e2119467119.
[52] Stern, A., Fleishon, S., Kustin, T., Mandelboim, M., Erster, O., Mendelson, E., Mor, O., and Zuckerman, N. S., 2021, "The unique evolutionary dynamics of the SARS-CoV-2 Delta variant," medRxiv.
[53] Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R. A., 2018, "Nextstrain: real-time tracking of pathogen evolution," Bioinformatics, 34(23), pp. 4121-4123.
[54] Du, P., Ding, N., Li, J., Zhang, F., Wang, Q., Chen, Z., Song, C., Han, K., Xie, W., Liu, J., Wang, L., Wei, L., Ma, S., Hua, M., Yu, F., Wang, W., An, K., Chen, J., Liu, H., Gao, G., Wang, S., Huang, Y., Wu, A. R., Wang, J., Liu, D., Zeng, H., and Chen, C., 2020, "Genomic surveillance of COVID-19 cases in Beijing," Nat Commun, 11(1), p. 5503.
[55] Wang, R., Chen, J., Gao, K., Hozumi, Y., Yin, C., and Wei, G. W., 2021, "Author Correction: Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants," Commun Biol, 4(1), p. 311.
[56] Hoffmann, M., Arora, P., Groß, R., Seidel, A., Hörnich, B. F., Hahn, A. S., Krüger, N., Graichen, L., Hofmann-Winkler, H., and Kempf, A., 2021, "SARS-CoV-2 variants B. 1.351 and P. 1 escape from neutralizing antibodies," Cell, 184(9), pp. 2384-2393. e2312.
[57] Ozer, E. A., Simons, L. M., Adewumi, O. M., Fowotade, A. A., Omoruyi, E. C., Adeniji, J. A., Olayinka, O. A., Dean, T. J., Zayas, J., and Bhimalli, P. P., 2022, "Multiple expansions of globally uncommon SARS-CoV-2 lineages in Nigeria," Nature communications, 13(1), pp. 1-13.
[58] Planas, D., Veyer, D., Baidaliuk, A., Staropoli, I., Guivel-Benhassine, F., Rajah, M. M., Planchais, C., Porrot, F., Robillard, N., Puech, J., Prot, M., Gallais, F., Gantner, P., Velay, A., Le Guen, J., Kassis-Chikhani, N., Edriss, D., Belec, L., Seve, A., Courtellemont, L., Péré, H., Hocqueloux, L., Fafi-Kremer, S., Prazuck, T., Mouquet, H., Bruel, T., Simon-Lorière, E., Rey, F. A., and Schwartz, O., 2021, "Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization," Nature.
[59] CDC, May 2022, "covid-data-tracker."
[60] Planas, D., Saunders, N., Maes, P., Guivel-Benhassine, F., Planchais, C., Buchrieser, J., Bolland, W.-H., Porrot, F., Staropoli, I., and Lemoine, F., 2022, "Considerable escape of SARS-CoV-2 Omicron to antibody neutralization," Nature, 602(7898), pp. 671-675.
[61] McCarthy, K. R., Rennick, L. J., Nambulli, S., Robinson-McCarthy, L. R., Bain, W. G., Haidar, G., and Duprex, W. P., 2021, "Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape," Science, 371(6534), pp. 1139-1142.
[62] Martin, D. P., Weaver, S., Tegally, H., San, J. E., Shank, S. D., Wilkinson, E., Lucaci, A. G., Giandhari, J., Naidoo, S., and Pillay, Y., 2021, "The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages," Cell, 184(20), pp. 5189-5200. e5187.

 doi:10.20944/preprints202210.0241.v1

[63] Lei, X., Dong, X., Ma, R., Wang, W., Xiao, X., Tian, Z., Wang, C., Wang, Y., Li, L., Ren, L., Guo, F., Zhao, Z., Zhou, Z., Xiang, Z., and Wang, J., 2020, "Activation and evasion of type I interferon responses by SARS-CoV-2," Nat Commun, 11(1), p. 3810.
[64] Liang, T., Cheng, M., Teng, F., Wang, H., Deng, Y., Zhang, J., Qin, C., Guo, S., Zhao, H., and Yu, X., 2021, "Proteome-wide epitope mapping identifies a resource of antibodies for SARS-CoV-2 detection and neutralization," Signal transduction and targeted therapy, 6(1), pp. 1-3.
[65] Smith, C. C., Olsen, K. S., Gentry, K. M., Sambade, M., Beck, W., Garness, J., Entwistle, S., Willis, C., Vensko, S., and Woods, A., 2021, "Landscape and selection of vaccine epitopes in SARS-CoV-2," Genome medicine, 13(1), pp. 1-23.
[66] Mou, K., Mukhtar, F., Khan, M. T., Darwish, D. B., Peng, S., Muhammad, S., Al-Sehemi, A. G., and Wei, D.-Q., 2021, "Emerging Mutations in Nsp1 of SARS-CoV-2 and Their Effect on the Structural Stability," Pathogens, 10(10), p. 1285.
[67] Lin, J. W., Tang, C., Wei, H. C., Du, B., Chen, C., Wang, M., Zhou, Y., Yu, M. X., Cheng, L., Kuivanen, S., Ogando, N. S., Levanov, L., Zhao, Y., Li, C. L., Zhou, R., Li, Z., Zhang, Y., Sun, K., Wang, C., Chen, L., Xiao, X., Zheng, X., Chen, S. S., Zhou, Z., Yang, R., Zhang, D., Xu, M., Song, J., Wang, D., Li, Y., Lei, S., Zeng, W., Yang, Q., He, P., Zhang, Y., Zhou, L., Cao, L., Luo, F., Liu, H., Wang, L., Ye, F., Zhang, M., Li, M., Fan, W., Li, X., Li, K., Ke, B., Xu, J., Yang, H., He, S., Pan, M., Yan, Y., Zha, Y., Jiang, L., Yu, C., Liu, Y., Xu, Z., Li, Q., Jiang, Y., Sun, J., Hong, W., Wei, H., Lu, G., Vapalahti, O., Luo, Y., Wei, Y., Connor, T., Tan, W., Snijder, E. J., Smura, T., Li, W., Geng, J., Ying, B., and Chen, L., 2021, "Genomic monitoring of SARS-CoV-2 uncovers an Nsp1 deletion variant that modulates type I interferon response," Cell Host Microbe, 29(3), pp. 489-502 e488.
[68] Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B., 2019, "The immune epitope database (IEDB): 2018 update," Nucleic acids research, 47(D1), pp. D339-D343.