Investigating the effect of random single-nucleotide polymorphisms (SNPs) on COVID-19 variant fitness to forecast epidemiological outcomes from probable mutated variants of concerns in South Africa.

How can the epidemiological properties of different COVID-19 epidemics be understood from trends in genetic mutations in the spike protein domain and their direct contribution to a variant's fitness and dominance?

Author: Dev Patel

Topic

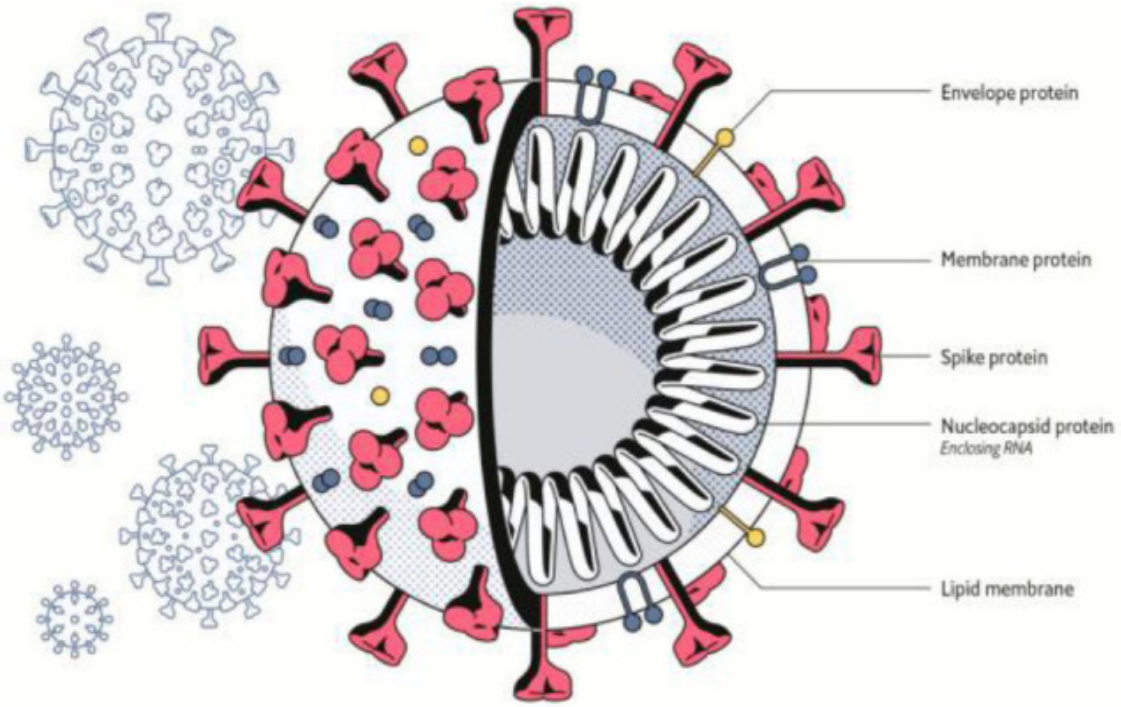
Health and development: the application of genomic surveillance in understanding and curtailing COVID-19 viral strains and their epidemiological consequences at a policy and biological level.

Approach

This paper examines the question: to what extent can bayesian regression forecast the epidemiological properties of different COVID-19 epidemics in order to predict and identify probable genetic mutations in the spike protein domain that contribute to a variant's fitness, dominance, and the effectiveness of public health response in South Africa across different waves. The vast majority of policies surrounding mask mandates, isolation, rapid-antigen testing, and community measures have constantly adapted and in many ways failed to curb the spread of arising novel variants of concern (VoC). While the vast majority of viral mutations are expected to be either deleterious or inherently neutral, certain missense SNPs (amino acid changes) can affect the functional properties of key proteins like the spike protein, which is the target for global vaccine formulations (Receptor-Binding Sites) and influences the behavior of the virus (infectivity, severity, and interactions with host immunity).

Introduction:

Pathogenic viruses are the cause of the vast majority of deadly disease outbreaks due to their ability to accumulate heritable genetic changes as a direct result of adapting to stressors in their environment. This not only makes it difficult to reduce the spread of pandemics (because of the heterogeneity of cases at a regional level), but also allows viruses to evade current



monitoring and diagnostic attempts. This is an acute contribution to the longevity and severity of the COVID-19 epidemic in regions that face these often unknown evolutionary pressures. This has made antigen testing, vaccines, and public health responses ineffective because they are inherently sequence-based tools that fail to work if the genome of these variants changes. As such, government institutions and local municipalities are forced to respond to new variants frequently and deal with the spreads as individual contagions, thereby limiting the use of preventative measures such as lockdowns and herd-immunization. All of these issues stem from the biological mechanisms that allow viruses to have emergent properties that help them develop resistance against our approaches to reducing the spread of disease.

In order to effectively mitigate the spread of COVID-19, each outbreak needs to be surveyed and understood at a local level from the regional mutations, circulation patterns, symptoms and vulnerable demographics, along with evaluating emergent variants as early as possible. However, conventional methods of monitoring new variants require these new mutated strains to be discovered first, followed by the whole-genome sequence of the patient and further testing and analysis of the mutations they have. This is a time-consuming and often expensive process as we are unable to make preventative changes in managing the spread of disease and are forced to be reactive in our measures. As genome sequencing technology and public datasets continue to increase in accessibility and speed, genomic surveillance can be a powerful tool in detecting new variants, but there is still a gap in being able to predict the most likely impact of these variants and their relative mutations when they are presented.

As such, this investigation aims to understand and evaluate these variants' properties in time and extrapolate them out across different patient cohorts to then identify the risk levels they pose. The goal of this inquiry is to look at a large set of SARS-CoV2 variants, and predict what kind of mutations are the most likely to occur in the Spike Protein region of the virus. A Bayesian regression model can be used to model the fitness of these mutations across 3 primary criteria: growth rate of lineages as a function of transmissibility and incidence, pathogenic properties of the virus based on endemicity, and the mutation rates of these variants. Mapping these possible mutations to phenotypes and properties will be essential for public health officials in making relevant containment decisions on factors such as virality, transmissibility, immunogenicity, death-rate, symptoms, rate of infection, and more. As the meta-data for these lineages includes geographies, the incidence and pattern of mutations in different countries can be used to plan out custom approaches to pandemic prevention.

Background:

Coronaviruses are RNA-based viruses that are pathogens to humans, causing Severe acute respiratory syndrome. The genome consists of 4 essential structural genes: membrane-glycoprotein-M (cellular communication and signaling), spike-protein-S (surface-proteins that transfer genetic material into cells), small-membrane-protein-SM, and a nucleocapsid-protein-N.

- S: large anchor attachment to the receptor including viral binding and host cell boundary by inserting into the endoplasmic reticulum
- M: virion envelope structural component, exchange of materials in and out of virus
- SM: miniature polypeptide, measuring 76–109 AAs, negligible component of virus
- N: measuring between 401-429 AAs, phosphoprotein part of the helical nucleocapsid that is believed to join the genomic RNA in a beads-on-a-thread manner

The SARS-CoV-2 virus is responsible for the COVID-19 pandemic, and has the greatest basic reproduction number (R_o) of any member of the Coronavirus family. With just a 70% genetic overlap with SARS-CoV-1, COVID-19 has a higher multiplication rate allowing it to be spread more efficiently and increase the rate at which the virus can evolve. Furthermore, its angiotensin-converting enzyme (ACE-2) receptors make it more efficient at avoiding host cells due to its high variation, preventing the body's immune cells from recognizing the pathogen. While the vast majority of transmissions take place in a short contact range, virus transmission can follow patterns of contact, ambient surroundings, host infectivity, and socio-economic factors that influence the risk of transmission.

Global and Local Scope:

Importance of Genomic Surveillance and Biological Impact on Virus Evolution

The greatest source of the economic, health, and social burden caused by the COVID-19 pandemic is the prevalence of the genetic and geographic variations of SARS-CoV-2. To evaluate the relative impact of new variants and develop specific plans of target, genomic surveillance is essential to manage and track the spread and evolutionary changes in these new strains and to evaluate their impact on a country's demographic and specific patient population properties. The greater problem countries that are more vulnerable to COVID-19 and its VoCs face due to them being a point of origin or not having effective vaccine programs to develop herd immunity is based on understanding a variant's impact on immune evasion and viral transmissibility. This is because simulating and matching variants to specific phenotypes and modeling viral behavior is limited by the available genetic data a country has.

Variants of Concern have been the primary cause for South Africa's recurring waves of disease spread, each with different kinds of transmission dynamics. The country serves as a useful case study when understanding the mutability and epidemiological properties of SARS-CoV-2

as the most heavily mutated strain Omicron originated from the nation and became the epicenter for a new international wave of infections. South Africa abnormally experienced 4 unique waves of COVID-19 caused by every major VoC identified throughout the pandemic, and each subsequent variant had a higher transmissibility and infection rate than the last. Not only did Omicron initially surge in the country, but the Beta variant (B.1.351) was especially dominant and first identified in late 2020 in South Africa, sparking the second global wave. The evolutionary dominance and battle between these variants was difficult to understand because only certain regions were being impacted at first by these new variants. Because South Africa stood out as an outlier having the home-grown variant advantage 2 out of the 4 times, genomic surveillance began to play a more pronounced role in understanding these properties for the respective variants.

South Africa Health Policy:

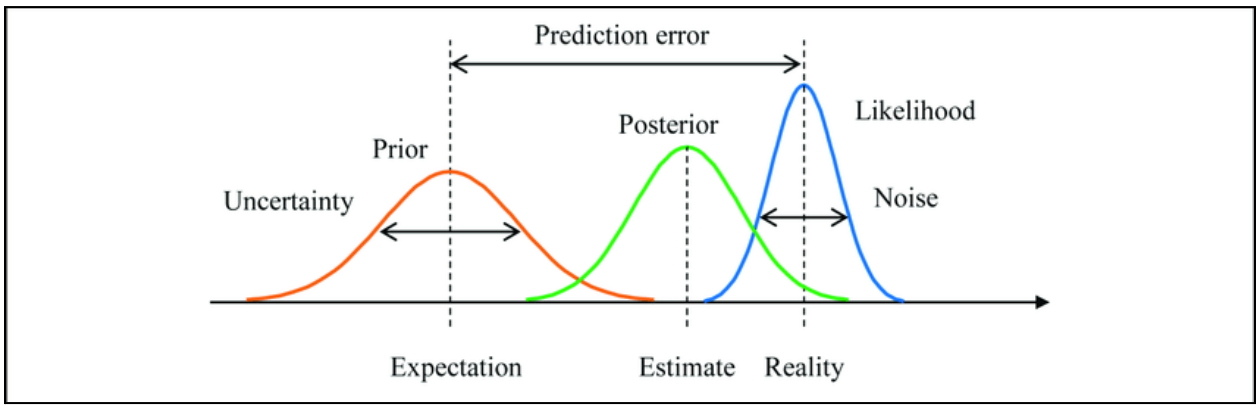
The country reported their first 100 cases on March 18, 2020. They have one of the highest Global Health Security Index Scores (54.8) across Sub-Saharan Africa and additionally have a high Epidemic PReparedness Index Score of 71.5, again making them more prepared than others Nevertheless, the country has an extremely high Fragile State index score, ranking 85th with a score of 70.1. With a population of 58,558,270, the vast majority is concentrated around the southern and southeastern coast far more than the western population. Nevertheless, South Africa's relatively high socioeconomic indicators are disparate from other essential indexes. For instance, their economic freedom score indicates that the majority of citizens are mostly unfree with one of the lowest health expenditure rates in the continent (8.1%). The vast majority of the younger (15-65) population is subject to countless diseases. More than 26% of the population is affected by mortality from cardiovascular diseases, cancer, diabetes, or chronic respiratory diseases. Additionally, the country has the highest incidence of HIV in the continent (19%). Nevertheless, they have a high Immunization coverage rate of 84%, As such, having a relatively high immunocompromised population puts the country at a greater risk with limited vaccine efficacy. The Beta Mutation and C.1.2 mutations originating from South Africa puts it in a unique position for being the epicenter for the highest mortality rates amongst any other countries in the region. South Africa made up 57.5% of the total confirmed cases back in September of 2020 out of the total 1.153 million confirmed cases at the time. It has a poor health system, coupled with a fragile economy, which led to strategies that efficiently make use of limited resources in managing exponential increases in outbreak of disease.

The key strategies that would benefit South Africa the most would be to allocate resources based on prevalent genes and mutations in geographic settings to provide "targeted interventions" (effective drugs, vaccine deployment, etc.).

Computational Methods Review:

Model-Inference Systems Epidemiological Properties for Transmission Dynamics

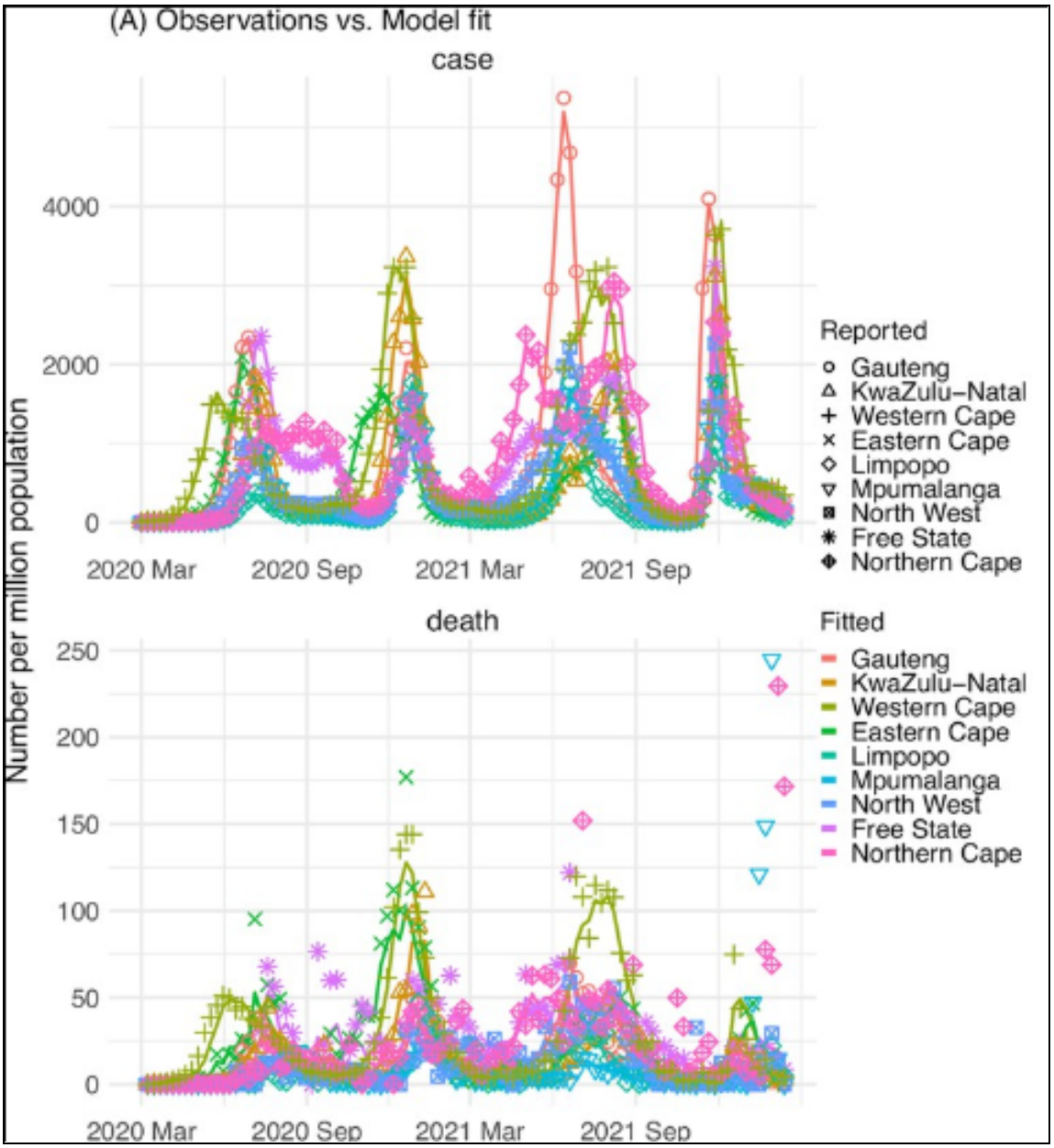
Epidemiologists can use individual case data, population dynamics parameters, and mortality rates to develop a deeper understanding of a variant's transmission dynamics. These are fit to the model using bayesian inference, which relies on weekly data to train a likelihood function around the prior and posterior. Bayes theorem states that the probability of an event is the proportion of said event occurring among all possible outcomes. This is modeled by a joint distribution of the prior and the likelihood, where the prior represents a uniform distribution that models the probability of all random variables occurring while the likelihood is a curve representing the observed values in a dataset. The joint distribution models a posterior, which acts as the expected value given a set of inputs, can help identify connections across latent and observable variables to predict useful information about our parameters and their correlations.



$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

The likelihood function models multivariate relationships by training a normal distribution P(B) on a set of observed data that is considered the ground truth P(A). By sampling an established dataset, the likelihood function simply maps a posterior/defined probability distribution to the prior.

These models are inherently useful because as the sample size of the likelihood/observable data increases, the posterior most closely models the density formula of the model. This is referred to as the objective function, and the bayesian model is meant to approximate this arbitrary function by iterating across different data points, thus changing the prior using a cost function. The reason for using Bayes is because it can ideally change around a set of multiple means to better reflect the different waves and variations in the data.



Source

For instance here, the posterior maps the normal distribution to observed data that has been reported in several provinces in South Africa for a total number of cases per million population. This can be repeated for other studies like serology data to calculate underlying infection rates. Yang et al were able to calculate changes in seroprevalence across multiple timestamps to observe changes in infection rates as a result of antibody counts and reinfection rates to better reflect mutations and dominant variants. The model was also used to correlate infection numbers with hospitalizations for each wave. Results from these inference models showed that the lower cases per capita were correlated to high under infection. In fact, the rate at

which infections were detected went down significantly as new waves were introduced despite having greater transmissibility.

A common epidemic model used by epidemiologists that leverages a Bayesian approach essentially analyzes the derivative of complex likelihood functions to make predictions on susceptible population infection rates.

$$\frac{dS}{dt} = \frac{R}{L_t} - \frac{b_t e_t m_t \beta_t IS}{N} - \varepsilon - \nu_{1,t} - \nu_{2,t}$$

$$\frac{dE}{dt} = \frac{b_t e_t m_t \beta_t IS}{N} - \frac{E}{Z_t} + \varepsilon$$

$$\frac{dI}{dt} = \frac{E}{Z_t} - \frac{I}{D_t}$$

$$\frac{dR}{dt} = \frac{I}{D_t} - \frac{R}{L_t} + \nu_{1,t} + \nu_{2,t}$$

Yang et al. S = total susceptible population (stratified by age and demographics), E = exposed population, I = Infected population, R = recovered population, N = total population size, epsilon = total travel-imported infections (this was negligible for Omicron and Beta while much significant for Delta and Alpha).

The first derivative looks at the total change in susceptible population as the recovered population / total immunity period L_t (rate at which population loses antibody protection) - the % of the population that was predicted to be infected at transmission rate β. b_t, e_t, and m_t are the Bayesian inference model's initialization parameters (season-scaled infection rate (0 if summer → 1 if winter to reflect increased conductive infection seasonality rates, NPI effectiveness as total interventions deployed/N population, mobility of population respectively). The vaccination rates are modeled as the total population vaccinated with a first dose (v_1) and second dose (v_2) multiplied by the relative vaccine effectiveness for each dose. South Africa had approximately ⅔ of its population immunized with a relative VE (VoC: v1, v2) of {Beta: 20/85%, Delta: 35/75%, Omicron: 10/35%). These model variables were estimated using the Bayesian model, and the derivatives were transformed into these epidemiological-modeling formulas to estimate changes in transmissibility and immune erosion for each respective variant.

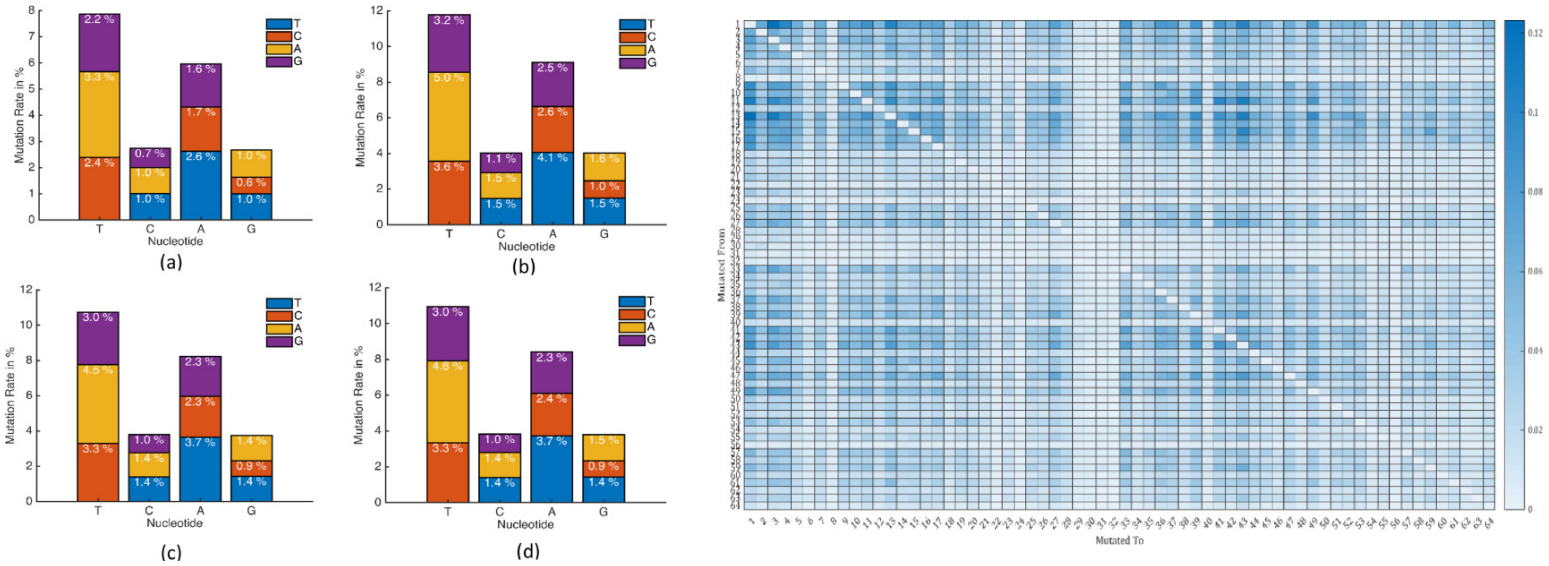
Quantity	Beta	Delta	Omicron (BA.1)
% Transmissibility Increase	34.3 (20.5, 48.2)	47.5 (28.4, 69.4)	94 (73.5, 121.5)
% Immune erosion	63.4 (45, 77.9)	24.5 (0, 53.2)	54.1 (35.8, 70.1)
% Infection detection	6.28 (σ = 0.8)	5.79 (σ = 1.0)	3.56 (σ = 0.5)
% Attack Rate	2.0	5.0	4.0

Homoplasies Frequency Count:

A homoplasy is a shared characteristic in a phylogenetic relationship that does not stem from a direct ancestor. By counting the presence or lack of these shared mutations at the single base-pair level, epidemiologists can look at the independent mutations across lineages and easily compare different variants amongst geographic areas as well. For instance, an allele frequency can give more direct population data for given countries and their biological significance can be used to explain the specific nature of the pandemic. Commonly, scientists use non-synonymous (does not lead to a change in the protein's amino-acid sequence, and thus does not change its function) and synonymous mutations (opposing, creates an isomer of the protein) and their substitution rates relative to a reference genome to calculate mutation rates and then predict the possibility of mutations occurring with this value. Furthermore, the distance between a reference genome and a variant genome offers an input to possible fitness models (discussed later).

The *nucleotide mutation rate* is defined as $MutationRate = \frac{\#mutations/distance}{lg*gs}$ where lg is the total number of sequences available and gs is the length of the Wuhan-Alpha-1 reference genome. This provides a matrix/table where the rows are specific patients (i) and columns are the nucleotides in the genome sequence (j).

As such, the mutation frequency for a given nucleotide is found at (i, j) and the mutation frequency itself is defined by $\sum_{i=0}^{n=lg} \sum_{j=0}^{n=gs} TABLE(i, j)$. $lg * gs$ provides the total area of the possible mutations. By indexing each mutation to a set of all possible nucleotides [A, T, C, G], a 4x4 lookup table is used to generate graphs on mutation frequencies and pairs. For instance, A -> [T, C, G] mutation sets can have different values.



The same model can be used for 64 codons as well, but instead the look up table generated will be 64x64. These codon mutations are distinct across all geographic regions, as the TA and GC content differs relative to the codon numbers.

Fitness Models and Inclusion of Viral Properties

The evolutionary fitness of a variant can be defined by a mapping function between the virological dynamics and the epidemiological properties of a virus. Empirical studies found that most random mutations had a -19% reduction of avg. fitness relative to phylodynamic models that are commonly used in the field of epidemiology.

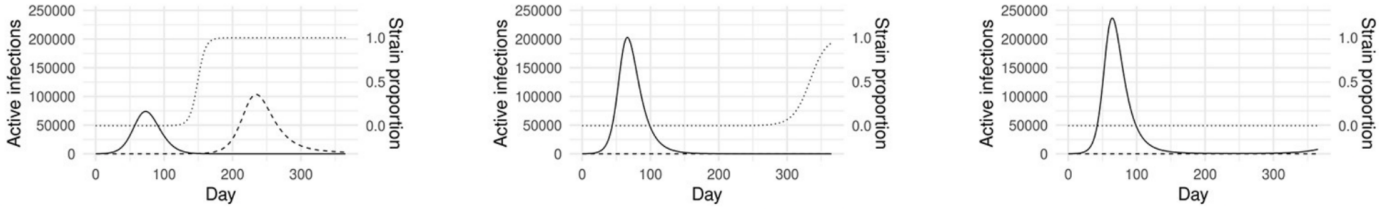


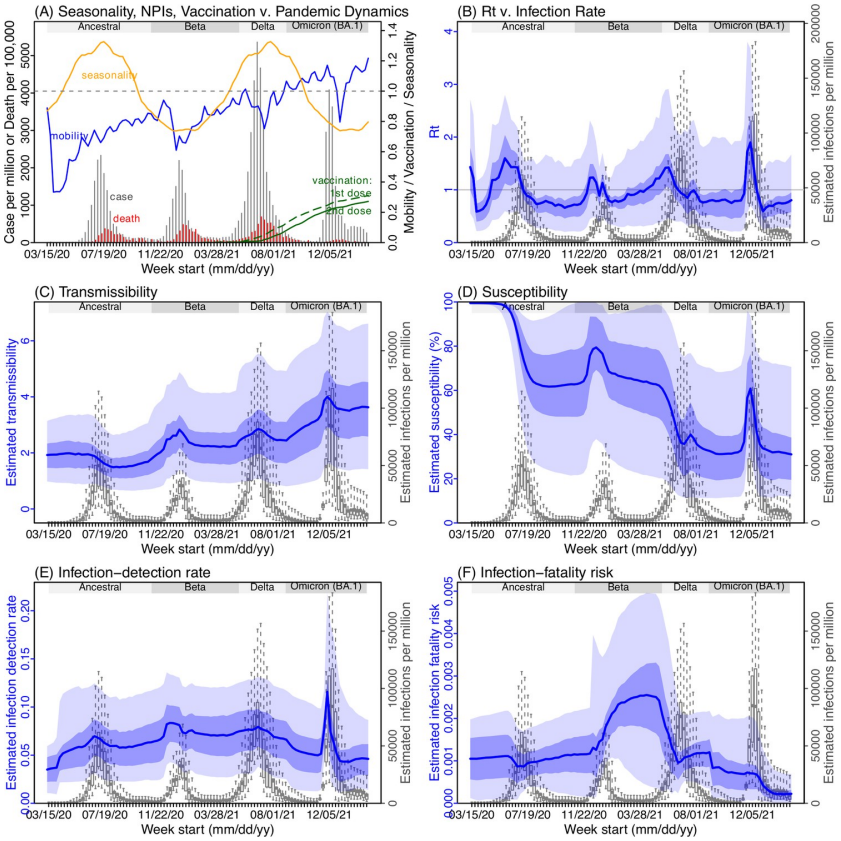
Fig 1. Example simulation results from a two-strain SIRS model with a vaccinated population, for different vaccination rates. The solid line shows the number of active infections for Strain 1. The dashed line shows the number of active infections for Strain 2. The dotted line shows the proportion of the two strains. Vaccine resistant strain initially present at rate of 10^{-6}. (**a**) 50% population vaccinated in 100 days. (**b**) 10% population vaccinated in 100 days. (**c**) No vaccination.

The most common models used in fitness are SIRS models, which model the rate of change in the total number of individuals who are susceptible, infected, resistant, or have lost immunity

over time as a function of the mutation rate of variants. Phylodynamic models can be used with these SIRS models to connect infection and transmissibility data with specific phenotypes. The fitness model as a result if modeled by the differential equation below that models the SIRS trend lines:

$$\frac{\partial}{\partial t}u = (x - \bar{x})u + \mu(u * g - u)$$

X is scalar quantity measuring the mutation fitness, u(x, t) is % of population with fitness x at a time t, u*g = integral u(x-R, t) g(R, t)dR, μ is mutation rate, x is integral of xudx which is average population fitness. The mutation likelihood is basically considered with a different shape and the fitness → the integrodifference equation in basically modeling a graph/fitness landscape aiming to model the properties that influence the fitness and its evolution.



Bayesian modeling from Yang et al was successfully able to generate clear correlations between a wide variety of disease dynamics. The primary reason for immune erosion from Beta to Omicron was not genetic but instead a higher population immunity amidst the country's identified susceptible population (as seen by the mismatch between spikes in infection rates

and the transmissibility with immune erosion).

The observed seasonality at any extreme greater than 50% of the mean line (blue) correlated directly with the total number of cases and vaccination efforts around the tailend. The reproduction number of Covid-19 spiked around in relation to the specific mutations that promoted higher infection rate spikes for every new Variant of Concern. The cumulative R_t and infection rate led to a consistent surge in transmissibility, which was found to not regress back to pre-pandemic levels due to later non-pharmaceutical and vaccine interventions being ineffective against newer strains. This backlog was discovered in tandem with the infection detection rate, where the lack of consistent genomic testing and surveillance failed to identify high incidences of variants relative to vaccination efforts. The relatively constant infection and R_t rate disguised the wave-specific abnormalities in transmissibility. The spike in infection-fatality risk occurred around the Beta VoC but quickly resided soon after Delta became the dominant VoC. The population susceptibility was found to be approximately 21.876% but Omicron's observed 44.9% attack rate was primarily attributed to reduced immunity from NCIs and ongoing vaccination efforts that fell flat until the tail-end of infections. However, it was interesting to see that the susceptibility decreased with each new VoC while transmissibility stayed constant. This could have been primarily attributed to the classification of vulnerable populations changing as a product of the growing mutation frequency in the spike protein over subsequent variants. This was primarily attributed to B.1.1.7/B.1.351's high population susceptibility and viral transmissibility increasing by 46.6% and 32.4% respectively.

Materials and Procedure:

Developing a toolset for predicting exactly where and which kinds of mutations are most likely to occur in the SARS-CoV-2 spike protein can help in predicting which health policies can dramatically improve the outcome of patients by observing the correlation between antigenic changes and their phenotypic advantages to the virus from an evolutionary fitness perspective. This inquiry evaluates the fitness of simulated viral strains using f(x) = reproduction rate + mortality rate and cross-correlating the fitness of a given mutation. Since each mutated proteome for a given SARS-CoV-2 strain is subjected to different evolutionary pressures, a reference genome (Wuhan-1) was used as a baseline for calculating the initial mutability rates. This was done via a Bayesian regression model, which used pre-existing NIH data to create a functional mutability landscape model that can predict mutable sites in the genome. The mutability site model was then applied to the South African genome to predict what kinds of mutations and where in the spike protein, ACE2 receptor, and RBD would occur, of which a set of possible recorded mutations with verifiable phenotypic properties were then matched with

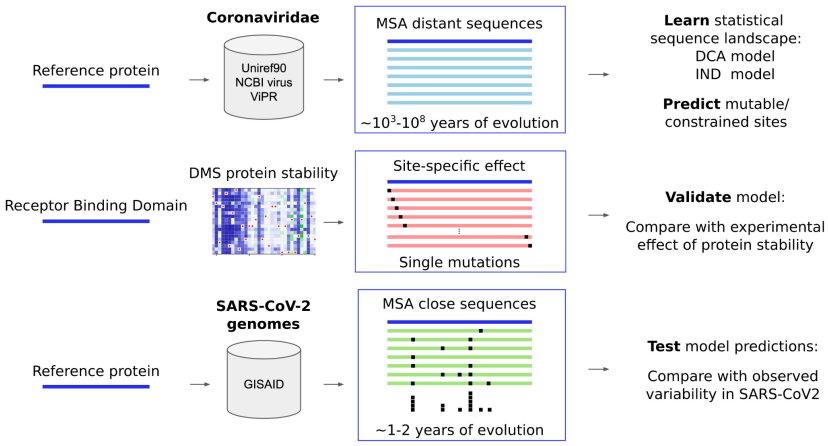
the strain's immunogenicity, transmissibility, and antigen immune-escape values. A reference chart was also developed to evaluate the health policies of South Africa and the spread of the virus to see which policies are directed at which epidemiological properties of the virus. This allowed for understanding what pandemic strategies from a health and policy standpoint directly target and whether said predicted mutations impact the effectiveness of South Africa's management of the pandemic.

After carefully evaluating all possible computational approaches to understanding the genetic makeup of mutations across variants in SARS-CoV-2, it became clear that the selected model had to be developed as a fitness model that could identify which mutations can increase infectivity or limit vaccine efficacy. This could be achieved by classifying mutations as variants of interests or concerns and then associating mutation frequency data with critical virus dynamic properties (transmissibility, efficacy, and antibody neutralization were the primary ones selected for exploration as they had a direct genotype-phenotype correlation over more epidemiological features that were already discovered via literature research). Furthermore, most mutated positions are heterogeneously distributed across COV-2 Proteins while the majority are on the spike protein, classified as either variations in other epitopes or binding sites or variations in therapeutic epitopes or binding sites (the spike protein S). These protein cores tend to be less variable as they can affect the stability of the protein fold and functionality. As such, exposed regions of spike protein have a large frequency of mutations resulting in variants with increased affinity to ACE2 receptor (primary mechanism of attack) and transmissibility. Understanding the specific amino acid residues would be the best place to look for the proteome as they are signs of selective pressure which affect the virus. Furthermore, using online datasets and clinical studies from NCBI and PubMed can offer insights on how peptide bonds and folding leaves exposed amino acids that then disassociate and act as residues that can be analyzed further.

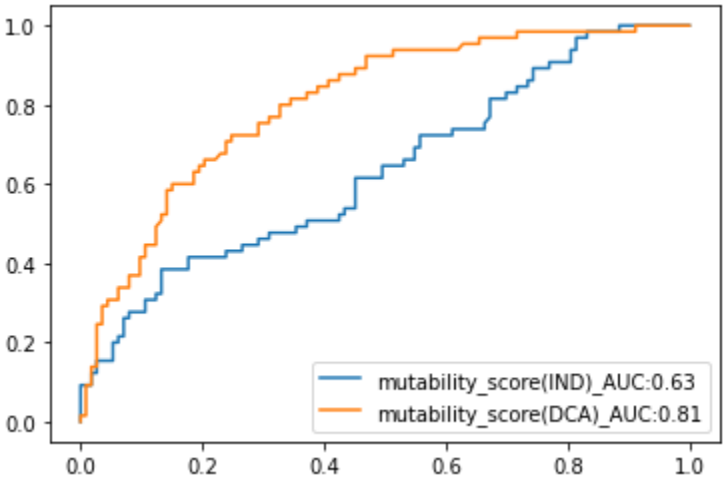
The models used in this study relied on the principle of epistasis which analyzes the dependence of mutation effects on other pre-existing mutations to cross-correlate patterns and residue mutations with their phenotypic effect. From the reference genome Alpha Wuhan-1, mutability scores could be generated by looking at the total observed mutations and properties like doubling rates and the kinds of mutations across different lineages from public genomic datasets (GISAID, Stanford NEXStrain, etc). To do this, I used Juan Rodriguez-Rivas et al's direct-coupling analysis data trained on over 1.2 million different conservation profiles. This is a statistical tool used to quantify the direct relationship between two positions and mutations in a biological sequence while simultaneously identifying how evolutionary pressures can affect networks of genes.

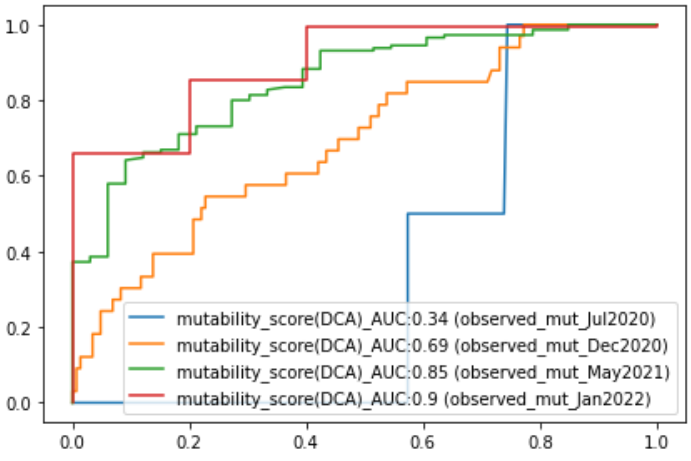
$$P(a|J,h) = \frac{1}{Z}\exp\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} J_{ij}(a_i,a_j) + \sum_{i=1}^{N} h_i(a_i)\right)$$

The DCA model can be applied for any set of overlapping genomic sequences across different variants and a reference to get a table of mutation frequencies and their respective nucleotide



classifications (as seen in the first methodology example that relies on multiple-sequence alignment pair distance).
Then, the model calculates mutability scores by cross-referencing these mutations with the receptor-binding domain's protein expression data from the Immune Epitope Dataset. After looking at all possible protein isomers that could be generated from a possible mutation, it maps the highest-frequency mutations circulating in a population to properties such as viral attachment, fusion, entry, and complementarity.



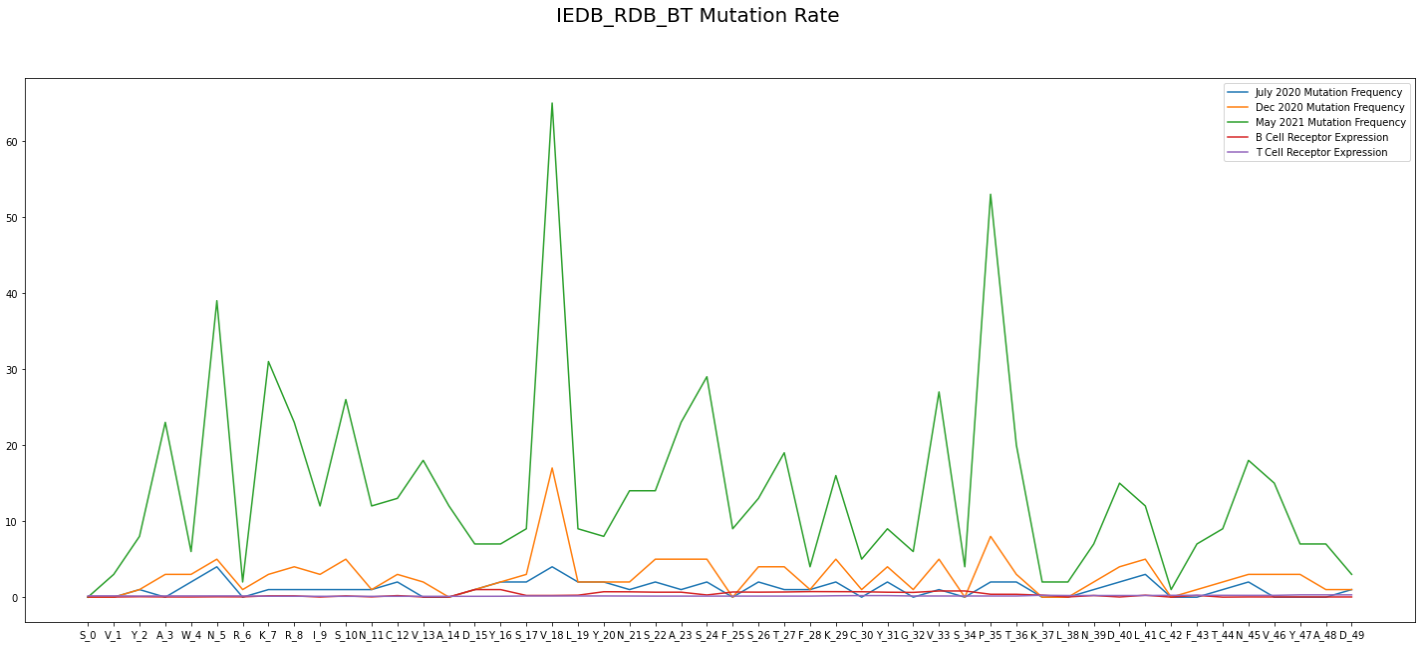


The DCA mutability score was calculated across 65 possible spike protein domains by observing the total distance between multiple variants of concerns and the Alpha-Wuhan-1 reference genome. Based on empirical data from GISAID, a site is classified as mutable if it has greater than 16 mutations. The rest of the 113 sites were classified as constrained, with the rare exception of previously thought to be highly conserved sites having some of the largest mutation frequencies but little diversity (see mutation table reference).

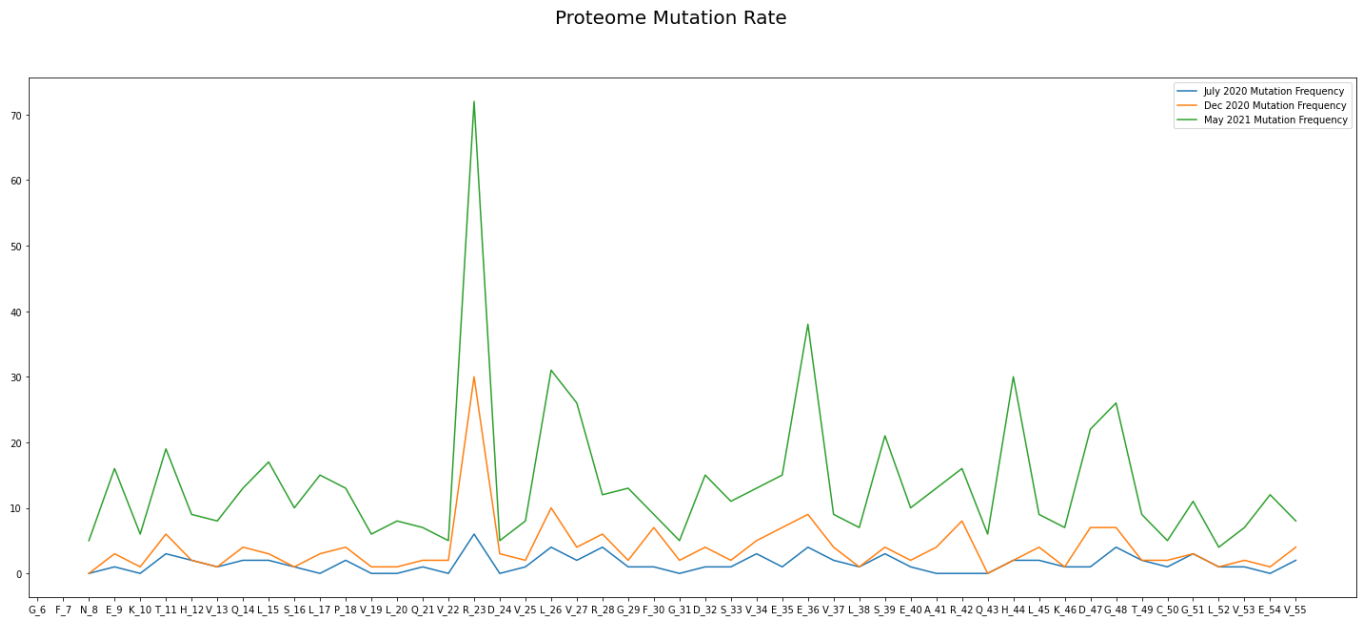
By taking the raw Immune Epitope Dataset and the Mutation Frequency table, I generated spike-domain and amino-acid specific mutability scores by observing the DCA mutability score across each amino acid position and its mismatch with the Wuhan-1 reference genome from July 2020 to January 2022. Moreover, the immune epitope data was cross-correlated with lower and upper-bound T-cell and B-cell expression rates.

```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt

# !git clone https://github.com/GiancarloCroce/DCA_SARS-CoV-2

ROOT = os.getcwd().replace('/main.ipynb', '/')
DATA = '/DCA_SARS-CoV-2/data/'

#Load Immune Epitope Database
meff_df = pd.read_csv(ROOT+DATA+'data_meff.csv', engine='python')
iedb_rdb_bt_df = pd.read_csv(ROOT+DATA+'data_dca_iedb_RDB_domain_BTcell.csv', engine='python')
iedb_rdb_df = pd.read_csv(ROOT+DATA+'data_dca_iedb_RDB_domain.csv', engine='python')
proteome_df = pd.read_csv(ROOT+DATA+'data_dca_proteome.csv', engine='python')


def make_graph(range_idx, df, title=None):
   fig, ax = plt.subplots(figsize=(25, 10))
   if title is not None:
       fig.suptitle(title, fontsize=20)
   x = [f"{b}_{a}" for a, b in enumerate(df['aa_Wuhan-Hu-1'].tolist())]
   y_may = df['observed_mut_May2021'].tolist()
   y_dec = df['observed_mut_Dec2020'].tolist()
   y_jul = df['observed_mut_Jul2020'].tolist()
   rf_bcell = df['rf_B_cell'].tolist()
   rf_tcell = df['rf_T_cell'].tolist()


   ax.plot(x[range_idx[0]:range_idx[1]], y_jul[range_idx[0]:range_idx[1]], label='July 2020 Mutation Frequency')
   ax.plot(x[range_idx[0]:range_idx[1]], y_dec[range_idx[0]:range_idx[1]], label='Dec 2020 Mutation Frequency')
   ax.plot(x[range_idx[0]:range_idx[1]], y_may[range_idx[0]:range_idx[1]], label='May 2021 Mutation Frequency')
   ax.plot(x[range_idx[0]:range_idx[1]], rf_bcell[range_idx[0]:range_idx[1]], label='B Cell Receptor Expression')
   ax.plot(x[range_idx[0]:range_idx[1]], rf_tcell[range_idx[0]:range_idx[1]], label='T Cell Receptor Expression')
```

```
    try:
        y_jan = df['observed_mut_Jan2022'].tolist()
        ax.plot(x[range_idx[0], range_idx[1]], y_may[range_idx[0], range_idx[1]], label='Jan 2022 Mutation
Frequency')
    except:
        pass

    ax.legend()

    plt.show()

make_graph([0, 50], iedb_rdb_bt_df, 'IEDB_RDB_BT Mutation Rate')
iedb_rdb_bt_df.head()
```



As shown for the first 49 codons on the graph, the most recent mutation rates were found to have seen magnitudinal spikes roughly around the same regions, with the biggest jump happening in May 2021 which connects back to the rapid Omicron spike in South Africa. Furthermore, B and T-cell expression levels remained relatively flat, indicating that almost all variants had some kind of mechanism for immune escape or antibody neutralization. The relative receptor binding domain mutation rates correlated with the total SARS-CoV-2 proteome for mutations outside of the spike.



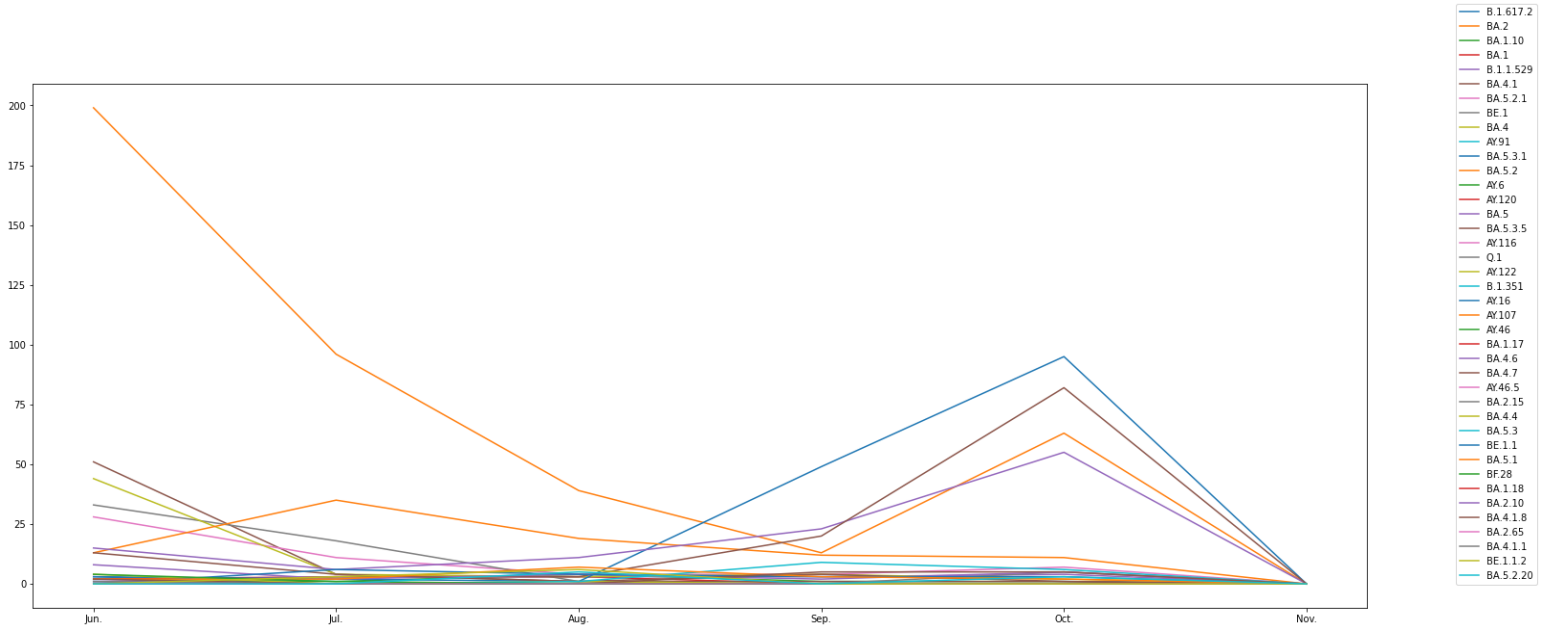
Most of these mutations were found in the nuclear envelope or ORF1a regions of the virus. After observing the top 40 spike mutations based on relative mutation frequency and epitope discrepancies, South Africa specific population data was collected on variant spread and paired with the specific mutation profiles generated for each given set of variants. These were collected from the NCBI variant tracker, and contained the direct clades, lineages, amino acid sequences and protein domains, nucleotide definitions, mutation types, and their relative frequencies.

lineage	synonyms	aa_definition	nt_definition	VinTEBS	CDC_status	NCATS_link	frequency
AY.107	B.1.617.2.107, Delta-AY.107	nsp4:V167L,S:P681R,S:H1101Y,M:I82T,N:T362I	G9053T,C23604G,C24863T,T26767C,C29358T	S:T478K, S:L452R	VBM	NaN	33
AY.116	B.1.617.2.116, Delta-AY.116	3CLpro:I213V,nsp13:P77L,M:I82T	A10691G,C16466T,T26767C	S:T478K, S:L452R	VBM	NaN	99
AY.120	B.1.617.2.120, Delta-AY.120	nsp3:P1228L,nsp3:S1230F,nsp13:P77L,S:S1147S	C6402T,C6408T,C16466T,A25003G	S:T478K, S:L452R	VBM	NaN	128
AY.122	B.1.617.2.122, Delta-AY.122	nsp2:K81N,nsp4:V167L,nsp13:P77L,M:I82T	G1048T,G9053T,C16466T,T26767C	S:T478K, S:L452R	VBM	NaN	69

AY.16	B.1.617.2.16,Delta-AY.16	nsp2:P129L,RdRp:G670S,ORF3a:S26L,M:I82T	C1191T,G15451A,C25469T,T26767C	S:T478K,S:L452R	VBM	NaN	46

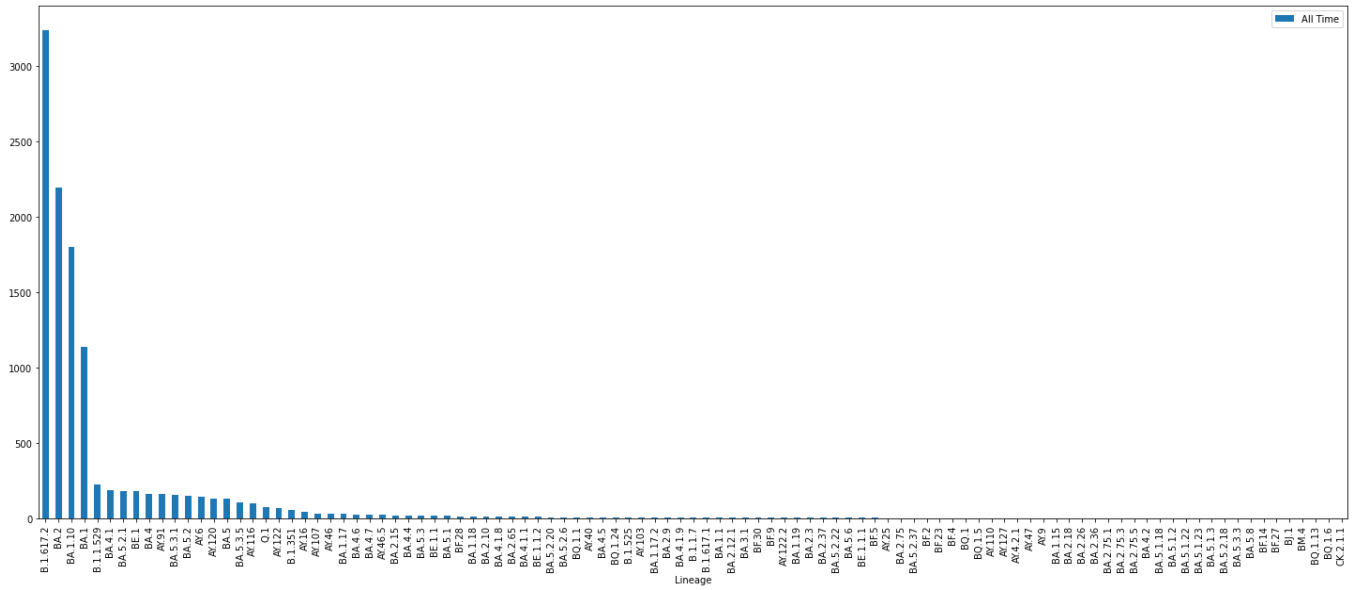
Specific epidemiological properties were collected from these respective variants and their cross-correlated mutations generated from the DCA coupling to produce trend lines for further analysis.

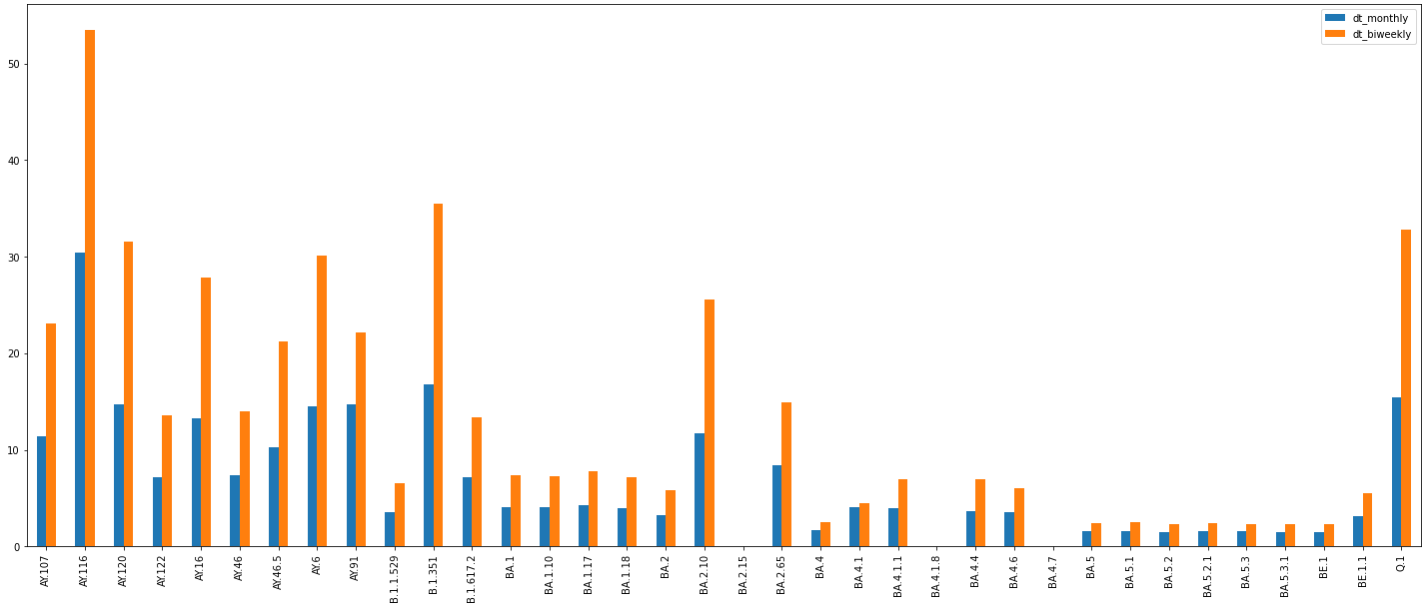
```
LINEAGES = '/lineages.csv'
lineages_df = pd.read_csv(ROOT+LINEAGES, engine='python')
sa_phylogeny = pd.read_csv(ROOT+'/sa-lineages.csv', engine='python')

sa_main = sa_phylogeny.iloc[:40]

legend = sa_main['Lineage'].tolist()
x = ['Jun.', 'Jul.', 'Aug.', 'Sep.', 'Oct.', 'Nov.']
y = np.asarray(sa_main[['Jun.', 'Jul.', 'Aug.', 'Sep.', 'Oct.', 'Nov.']])

fig, ax = plt.subplots(figsize=(25, 10))
for i in range(len(y)):
   ax.plot(x, y[i], label=legend[i])
fig.legend()
plt.show()

variant_data = lineages_df[lineages_df['lineage'].isin(legend)]
try:
   variant_data = variant_data.drop(columns=['doubling_time_mo_usa', 'doubling_time_bw_usa'])
except:
   pass

sa_phylogeny['All Time'] = sa_phylogeny['All Time'].replace(',', '', regex=True)
sa_phylogeny['All Time'] = pd.to_numeric(sa_phylogeny['All Time'])
sa_phylogeny.plot.bar(x='Lineage', y='All Time', figsize=(25, 10))

# variant_data.head()
variant_data['doubling_time_bw_not_usa'] = variant_data['doubling_time_bw_not_usa'].fillna(0)
variant_data['doubling_time_mo_not_usa'] = variant_data['doubling_time_mo_not_usa'].fillna(0)
index = variant_data['lineage'].tolist()

dt_mo = variant_data['doubling_time_mo_not_usa'].tolist()
dt_bw = variant_data['doubling_time_bw_not_usa'].tolist()
doubling_variant = pd.DataFrame({'dt_monthly': dt_mo, 'dt_biweekly': dt_bw}, index=index)
doubling_variant.plot.bar(figsize=(25, 10))

sa_main.head()
```



The top variants with their relative doubling period acting as a strong indicator for transmissibility and the total rate of infection (as specified in the methodology and design section) were then aggregated across the time series to generate a cumulative frequency of total incident cases to model the relative growth rates and spread of these variants plotted against their mutability scores.

The resulting set of variants that had the greatest growth rate and spread had their genetic sequences from the spike protein taken from GISAID public repositories and aligned against a FASTA reference file once again. From here, the specific mutations were classified from the nucleotide and amino-acid definitions provided in the reference lineages table.





The resulting set of variants that had the greatest growth rate and spread had their genetic sequences from the spike protein taken from GISAID public repositories and aligned against a FASTA reference file once again. From here, the specific mutations were classified from the nucleotide and amino-acid definitions provided in the reference lineages table.

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCT
GTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACT
CACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATC
TTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
CGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAAC
ACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGG
AGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACT
CGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGG
CGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGG
TGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
TCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGA
ACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTC
ATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCG
TGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCA
GACACCTTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAA
TTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAACCAAGGGTTGAAAAGAAAAA
GCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATG
CAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA
GACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGA

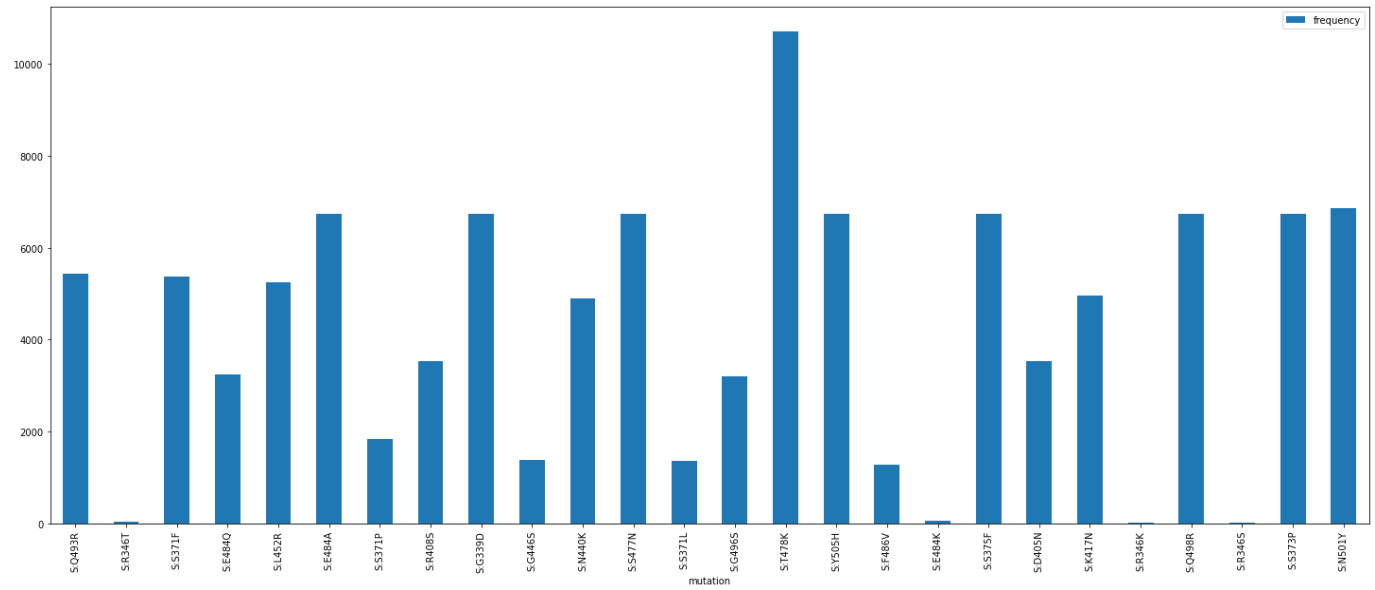
The first set of nucleotide pairs in the Wuhan-1 Reference Genome

```
from itertools import chain
```

```
unique_mutations = set(','.join(master_variant['VinTEBS'].unique()).split(','))
ref = pd.read_csv('gene_reference.csv', engine='python')
# print(unique_mutations)

total_freq = {key: [] for key in unique_mutations}
all_vintebs = master_variant['VinTEBS'].tolist()

for mut in list(unique_mutations):
   for idx, vintebs in enumerate(all_vintebs):
       if mut in vintebs:
           total_freq[mut].append(master_variant.iloc[idx]['frequency'])

mutation_frequency = {key: sum(value) for key, value in total_freq.items()}
print(mutation_frequency)


keys = mutation_frequency.keys()
values = mutation_frequency.values()

mut_freq = pd.DataFrame.from_dict({'mutation': list(keys), 'frequency': list(values)})

mut_freq.plot.bar(x='mutation', y='frequency', figsize=(25, 10))
mut_freq.iloc[:100]
```

The resulting unique mutations and their cumulative frequencies across the pandemic and its individual waves in South Africa were generated from this pipeline.

mutation	frequency
S:Q493R	5444
S:R346T	37
S:S371F	5369
S:E484Q	3237
S:L452R	5257
S:E484A	6732
S:S371P	1840
S:R408S	3529
S:G339D	6732
S:G446S	1376
S:N440K	4892
S:S477N	6732
S:S371L	1363
S:G496S	3203
S:T478K	10701
S:Y505H	6732
S:F486V	1288
S:E484K	53
S:S375F	6732
S:D405N	3529
S:K417N	4958
S:R346K	13
S:Q498R	6732
S:R346S	25
S:S373P	6732
S:N501Y	6862



After compiling the epitope phenotype data with the BT-Cell expression, variant mutation frequencies, and viral dynamic values across transmissibility, immune escape potential, and

replication, the following table was compiled to highlight the most relevant and significant mutation changes that occurred on the Alpha Wuhan-1 reference genome in South Africa across the spread of the pandemic.

Argument and Observation 1 from Epistatic Modelling: HIV and Immunocompromised Population

The primary reason for this is that South Africa, found from the epistatic modeling results, has the most immunocompromised populations in the world.

A study had a 36-year old woman with advanced HIV was tracked carrying the novel coronavirus for 216 days, and during this time the virus accumulated more than 30 mutations, 13 of which were made on the spike protein that have been identified to allow the virus to escape immune responses. There are also another 19 mutations that include the 241/242/243 deletion which change its behavior such that it can stay dormant without inducing symptoms for extended periods of time. HIV infections have been found to be a primary source of new variants because the patients could carry the virus for longer, thereby relying on immunosuppressed populations and their demographics to control the transmission and evolution of the virus. Other causes for immunosuppressed populations also include kidney transplants and persons with hematological malignancies or have undergone hematopoietic stem cell or solid organ transplants. Africa as a greater continent is the most subject to this. 20.6 million people out of its 37.6 million population live with HIV, and South Africa is disproportionately the country with the highest population of HIV patients. (Tulio de Oliveira, South African Universities Network). The country has over 8.2 million people infected with AIDs-causative viruses, and while vaccination rates and reach are relatively high, the majority of accessible antiretroviral medications do not work. This is why neighboring countries to South Africa, those of which that rely on the country's vaccine networks for internal pathogen management are further susceptible to the virus (Botswana, Zimbabwe, Eswatini).

Experimental results support the unique situation of South Africa and the argument around having a high immunocompromised population being a key driving factor for the high mutation and variation amongst Omicron and Delta strains along with their transmissibility. The accumulation of these mutations has led to greater pathogenic variant diversity and are validated by the mutation types that were the most common across South Africa and the world's variant genetic makeup.

The most common mutation found amongst all Omicron cases was the substitution N501Y in the spike protein, being found in 6862 from the total 11000 genomic samples. Commonly found in the B.1.1.7 Omicron variant, the spike protein non-synonymous mutation from Asparagine to Tyrosine created a substitution that increased transmission efficacy in patients

who were infected; these mutations acted as a tropism against increased viral competition from Beta and Alpha variants by manipulating the way in which the Spike protein interacts with cell receptors for infection. Out of 8 possible mutations and deletions with a similar non-synonymous profile, only N501Y enabled a direct fitness gain for replication and enhanced viral transmission by increasing the spike protein's affinity with cellular receptors in human epithelial cells. The mutation was first detected in South Africa, but was also found as a product of convergent evolution in Brazil indicating that this mutation has more to do with pressures that have led to selective prioritization of replicability and transmission to obtain evolutionary fitness. Via 8 screens, N501Y was found to have a substantial fitness benefit for the variant by staying in the body for longer periods of time directly in the nasal cavity while simultaneously increasing the viral load over time. As such, the direct phenotypic extremes and symptoms were not discovered with this mutation compared to the initial spike protein sequences found in the Alpha and Delta variants. This enabled the virus to stay in the trachea and lungs for longer periods of time while simultaneously staying undetected. In South Africa, preliminary population frequency studies found that more than 89.7% of all active COVID-18 cases around February 2021 were exposed to this mutation and directly correlated with the rampant spread of Omicron and the sudden decline of other competing variants. The substitution spread quickly amongst the South African population via Omicron and further increased spike protein binding affinity for the ACE2 receptor. The sudden increase in total cases and dormant emergent properties SARS-CoV-2 gained as a result from this mutation were primarily attributed to lower detection rates and limited testing, thereby skewing results on the magnitude of the wave and its origin in South Africa. In fact, recombinant receptor-binding domains with the N501Y substitution were found to have a 819-fold improvement in binding affinity to the ACE2 receptor, indicating a substantial increase in viral fitness on the basis of replication rates, enhanced transmissions, longer resting periods in the upper airway. [1]

Mutation S:R346S, while less common (25 individual instances) was found to facilitate the targeted escape of Omicron variants from precursor monoclonal antibodies produced by therapies like sotrovimab. Sotrovimab is another monoclonal antibody primarily used for immunocompromised patients and to inhibit VoC entry into the host cell. It achieves this by blocking the interaction between the spike protein's RBD and ACE2 binding site. While previously found to have a relatively conserved epitope in this region with little mutations, the substitution mutation was found to have increased the total number of contacts between the CDR region on the ACE2 receptor (determines complementarity of sequence and whether it can bind to receptor) and the secondary structure (alpha helix at the N-terminus on amino acids 337-344) to inhibit the effect of the mAb. This emergent property also allowed the Omicron variant to circumvent the inhibitory effect of imdevimab and casiriviamb that also rely on the high conserved epitopes on the receptor binding domain to execute their function.

Selecting for this mutation has been found to lead to targeted immune escape just 7 weeks after infection while increasing the virus' infection rate.[2]

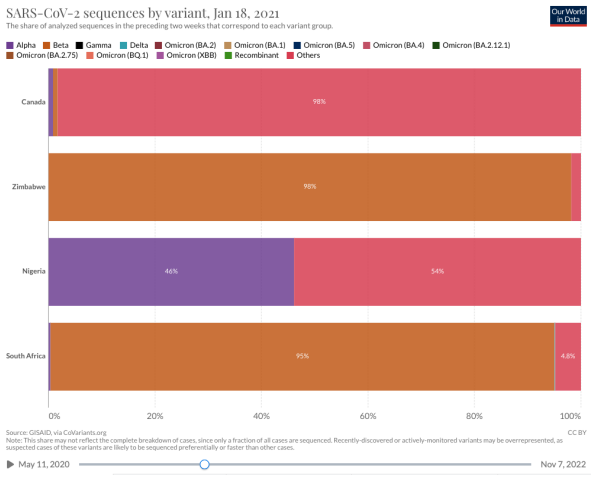
S:Q493 (glutamine to arginine), with a total frequency of 5444 cases, is a mutation found amongst people with mild to moderate COVID-19, first detected during cocktail drug trial studies for two monoclonal antibody treatments to treat high hospitalization and death rates for the Delta variant (bamlanivimab and etesevimab) but originated directly from Alpha infections. The Q493R mutation was identified during the treatment in a patient with cutaneous T-cell lymphoma who got infected with the VoC. The mutation was found in a patient with a compromised immunity, and the mutation itself allowed the virus to develop immune escape potential. The viral load following the detection of this mutation remained high even 40 days after the initial treatment. Its highest frequency was detected in the B.1.1.7 variant. The corresponding A → G substitution mutation led to a synonymous amino-acid substitution in the spike protein. Following the mutation, the binding activity of the spike protein increased dramatically, and it led to high levels of anti-spike binding antibody production in the body. The mutation was directly correlated in the lack of efficacy of the mAb, where the viral load change still left a total cycle threshold value of the PCR amplification reaction of 27.5, which indicated an increased presence of COVID-19 present markers. The mutation further corresponded to reduced susceptibility to antibody recognition, allowing the virus to escape the body's natural antibody receptor-binding response. The final recommendation from the original study team was to rapidly screen for the mutation in potential non-responders, particularly the immunocompromised population. Furthermore, this carried the idea that the T-cell to HIV to SARS-CoV-2 immune response provided the necessary evolutionary pressures that not only provided justification for stagnant B-T Cell activation levels (indicating that any new COVID-19 VoCs that became prominent were able to evade immune responses while those that could not reduced in total count) but solidified the reasoning behind South Africa being the origin point for 2 of the 4 variant waves (both of which possessed extremes of being pathogenic and transmissible). [3]

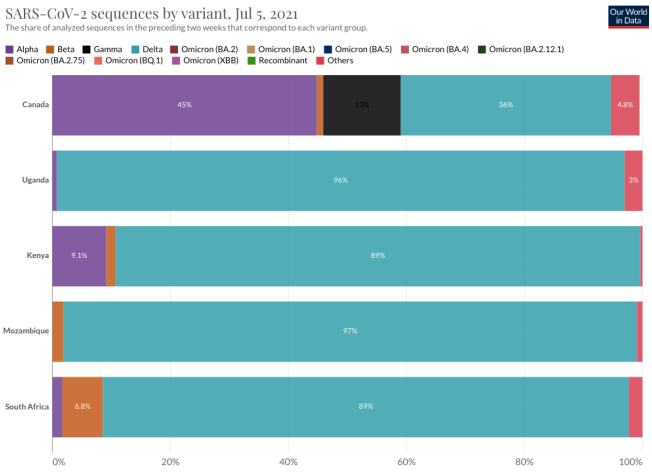
Out of the 30 mutations that are responsible for antibody escape from current vaccines and mAb therapies in Omicron, the G339D (6732) and N440K (4892) mutations were the only ones to have been found to significantly disrupt T-Cell responses. These mutations offer the phenotypic advantage of disrupting the body's natural or vaccine-induced immunity towards common Omicron variants. Mutations were found to have a down-regulating effect on CD memory cells. A t-test was conducted between four-week exposure to this variant vaccinated and unvaccinated at a p-value of 0.0008, and signified a major statistical difference between both patient groups' PBMC levels. The percentage of activated cells simulated by these peptide mutations was significantly reduced in Omicron variants that presented these mutations compared to those with the Alpha-variant mutations on the spike. Furthermore, G339D abnormally increased IFNy (peripheral blood mononuclear cell activity markers) while N440K

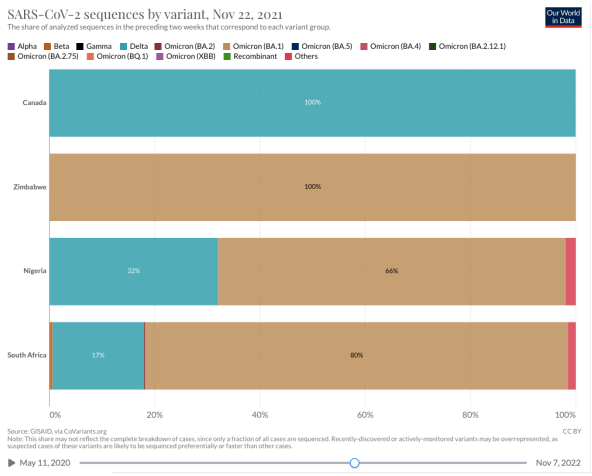
was largely responsible for its downregulation. The inverse relationship was found to have been a cause of further downregulating the more important AIM+ spike-specific T cells by weakening their memory on Omicron spike proteins, allowing the virus to evade the T-cell response completely within just 4 weeks of exposure to the patient. These are the most dangerous mutations since they occur in targets of human B and T cells, allowing the variant to quickly evade immune response induced by vaccination or previous infection. [9]

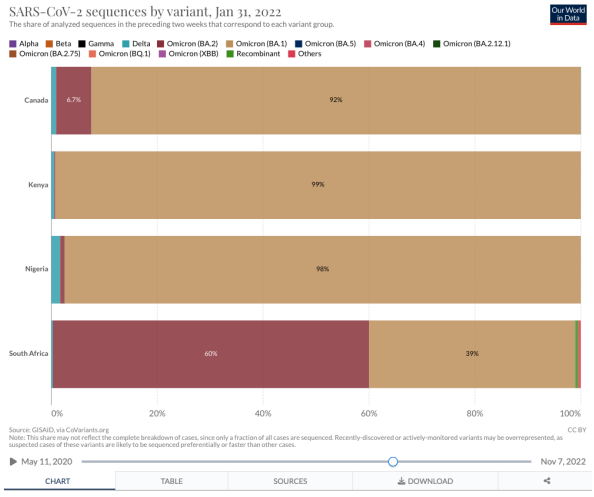
For defense against monoclonal antibody resistance (a common treatment administered in South Africa to combat HIV infections), the S:S371F mutation had the absolute highest prevalence across all Omicron BA, BQ, and XBB variants with a total frequency of 5369. The spike mutation is commonly paired with 5 additional spike protein mutations S:T19I, S:V213G, S:S371F, S:T376A, S:D405N, and S:R408S. While the mutation is a defining trait in all Omicron variants, it is still non-synonymous where the resulting protein is identical. The variant is commonly detected in sotrovimAB resistance, and was found to be found in all patients of 42 immunocompromised patients and those with severe asthma. The resulting spike protein mutation has 3 haplotypes, all of which lead to an increase in viral loads, particularly those who were fully vaccinated. The mutation was also detected in clinical studies with kidney transplant patients that underwent supportive oligonucleotide technique (SOT) treatment, where a specific genetic molecule is injected into a patient to control gene expression pathways responsible for replication. As such, this mutation in the spike protein is a leading cause for the high transmissibility rates that entail Omicron, while simultaneously originating from resistance from immunocompromised patients. [4]

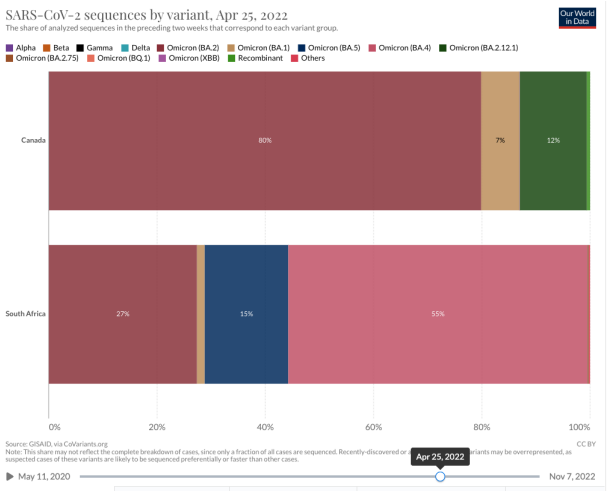
As a result, the rise of variants following the relative decline of Alpha, Beta, and Delta variants in other countries like the US and Canada could be a direct result of the lower vaccination rates. If we look at the cumulative variant share as a population of total COVID-19 cases, South Africa is the first to see substantial growth in the sequence diversity first and foremost, and these variants grow in spread at such a large rate that they soon become the dominant strain. Neighboring countries like Nigeria and Zimbabwe see the relative change in variant types before countries like Canada are quickly exposed to it later down the line. African nations, particularly South Africa, see the most dramatic and greatest shifts in variant types before other countries like Canada, the UK, and the US do.

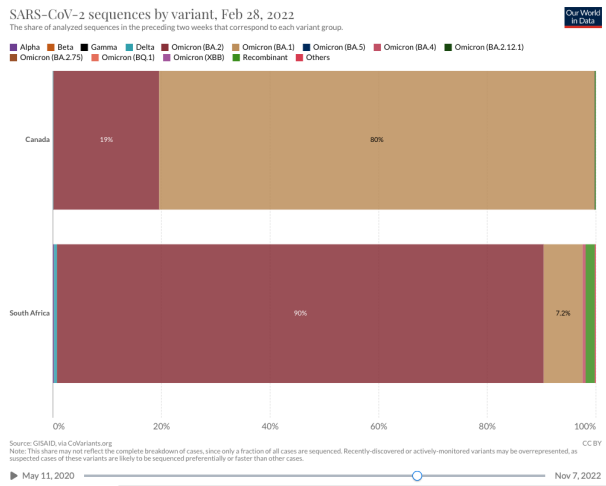

SARS-CoV-2 sequences by variant, Jan 18, 2021
The share of analyzed sequences in the preceding two weeks that correspond to each variant group.
Our World in Data
Alpha Beta Gamma Delta Omicron (BA.2) Omicron (BA.1) Omicron (BA.5) Omicron (BA.4) Omicron (BA.2.12.1)
Omicron (BA.2.75) Omicron (BQ.1) Omicron (XBB) Recombinant Others
Canada
Zimbabwe
Nigeria
South Africa
98%
98%
46%
54%
95%
4.8%
0%
20%
40%
60%
80%
100%
Source: GISAID, via CoVariants.org
Note: This share may not reflect the complete breakdown of cases, since only a fraction of all cases are sequenced. Recently-discovered or actively-monitored variants may be overrepresented, as suspected cases of these variants are likely to be sequenced preferentially or faster than other cases.
CC BY
May 11, 2020
Nov 7, 2022

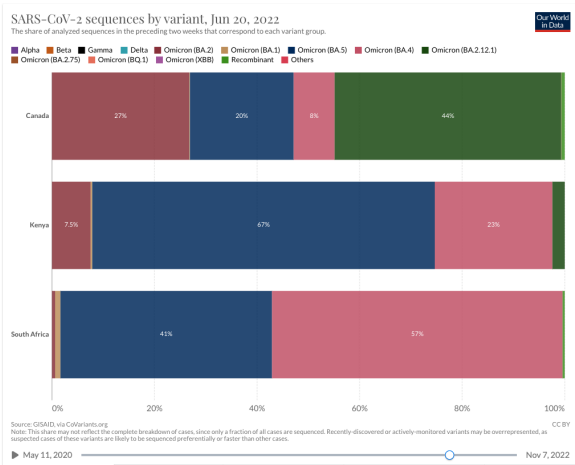

SARS-CoV-2 sequences by variant, Jul 5, 2021
The share of analyzed sequences in the preceding two weeks that correspond to each variant group.
Our World in Data
Alpha Beta Gamma Delta Omicron (BA.2) Omicron (BA.1) Omicron (BA.5) Omicron (BA.4) Omicron (BA.2.12.1)
Omicron (BA.2.75) Omicron (BQ.1) Omicron (XBB) Recombinant Others
Canada
Uganda
Kenya
Mozambique
South Africa
45%
13%
36%
4.8%
96%
3%
9.1%
89%
97%
6.8%
89%
0%
20%
40%
60%
80%
100%

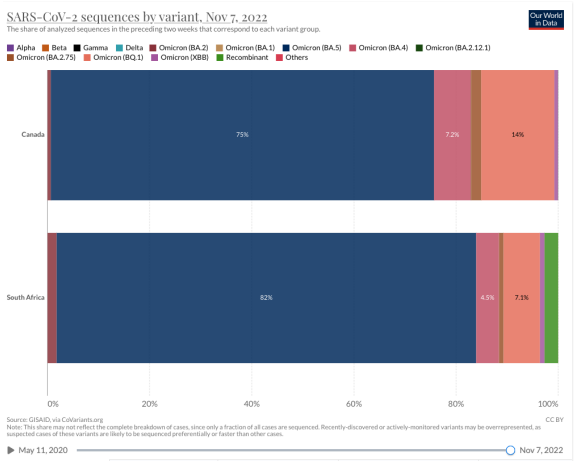

SARS-CoV-2 sequences by variant, Nov 22, 2021
The share of analyzed sequences in the preceding two weeks that correspond to each variant group.
Our World in Data
Alpha Beta Gamma Delta Omicron (BA.2) Omicron (BA.1) Omicron (BA.5) Omicron (BA.4) Omicron (BA.2.12.1)
Omicron (BA.2.75) Omicron (BQ.1) Omicron (XBB) Recombinant Others
Canada
Zimbabwe
Nigeria
South Africa
100%
100%
32%
66%
17%
80%
0%
20%
40%
60%
80%
100%
Source: GISAID, via CoVariants.org
Note: This share may not reflect the complete breakdown of cases, since only a fraction of all cases are sequenced. Recently-discovered or actively-monitored variants may be overrepresented, as suspected cases of these variants are likely to be sequenced preferentially or faster than other cases.
CC BY
May 11, 2020
Nov 7, 2022

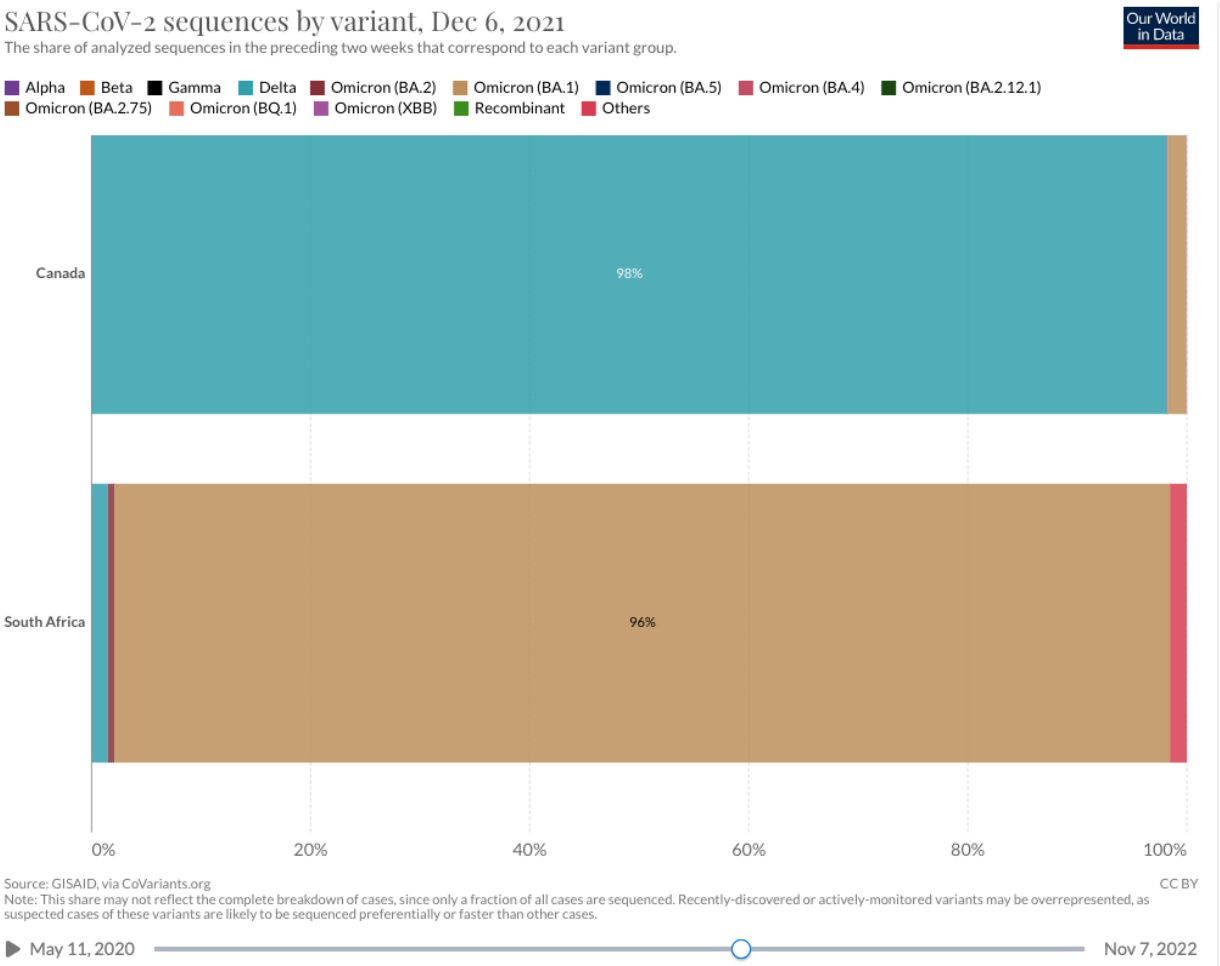

SARS-CoV-2 sequences by variant, Jan 31, 2022
The share of analyzed sequences in the preceding two weeks that correspond to each variant group.
Our World in Data
Alpha Beta Gamma Delta Omicron (BA.2) Omicron (BA.1) Omicron (BA.5) Omicron (BA.4) Omicron (BA.2.12.1)
Omicron (BA.2.75) Omicron (BQ.1) Omicron (XBB) Recombinant Others
Canada
Kenya
Nigeria
South Africa
6.7%
92%
99%
98%
60%
39%
0%
20%
40%
60%
80%
100%
Source: GISAID, via CoVariants.org
Note: This share may not reflect the complete breakdown of cases, since only a fraction of all cases are sequenced. Recently-discovered or actively-monitored variants may be overrepresented, as suspected cases of these variants are likely to be sequenced preferentially or faster than other cases.
CC BY
May 11, 2020
Nov 7, 2022
CHART
TABLE
SOURCES
DOWNLOAD


SARS-CoV-2 sequences by variant, Apr 25, 2022
The share of analyzed sequences in the preceding two weeks that correspond to each variant group.
Our World in Data
Alpha Beta Gamma Delta Omicron (BA.2) Omicron (BA.1) Omicron (BA.5) Omicron (BA.4) Omicron (BA.2.12.1)
Omicron (BA.2.75) Omicron (BQ.1) Omicron (XBB) Recombinant Others
Canada
South Africa
80%
7%
12%
27%
15%
55%
0%
20%
40%
60%
80%
100%
Source: GISAID, via CoVariants.org
Note: This share may not reflect the complete breakdown of cases, since only a fraction of all cases are sequenced. Recently-discovered or
variants may be overrepresented, as suspected cases of these variants are likely to be sequenced preferentially or faster than other cases.
CC BY
Apr 25, 2022
May 11, 2020
Nov 7, 2022


SARS-CoV-2 sequences by variant, Feb 28, 2022
The share of analyzed sequences in the preceding two weeks that correspond to each variant group.
Our World in Data
Alpha Beta Gamma Delta Omicron (BA.2) Omicron (BA.1) Omicron (BA.5) Omicron (BA.4) Omicron (BA.2.12.1)
Omicron (BA.2.75) Omicron (BQ.1) Omicron (XBB) Recombinant Others
Canada
South Africa
19%
80%
90%
7.2%
0%
20%
40%
60%
80%
100%
Source: GISAID, via CoVariants.org
Note: This share may not reflect the complete breakdown of cases, since only a fraction of all cases are sequenced. Recently-discovered or actively-monitored variants may be overrepresented, as suspected cases of these variants are likely to be sequenced preferentially or faster than other cases.
CC BY
May 11, 2020
Nov 7, 2022

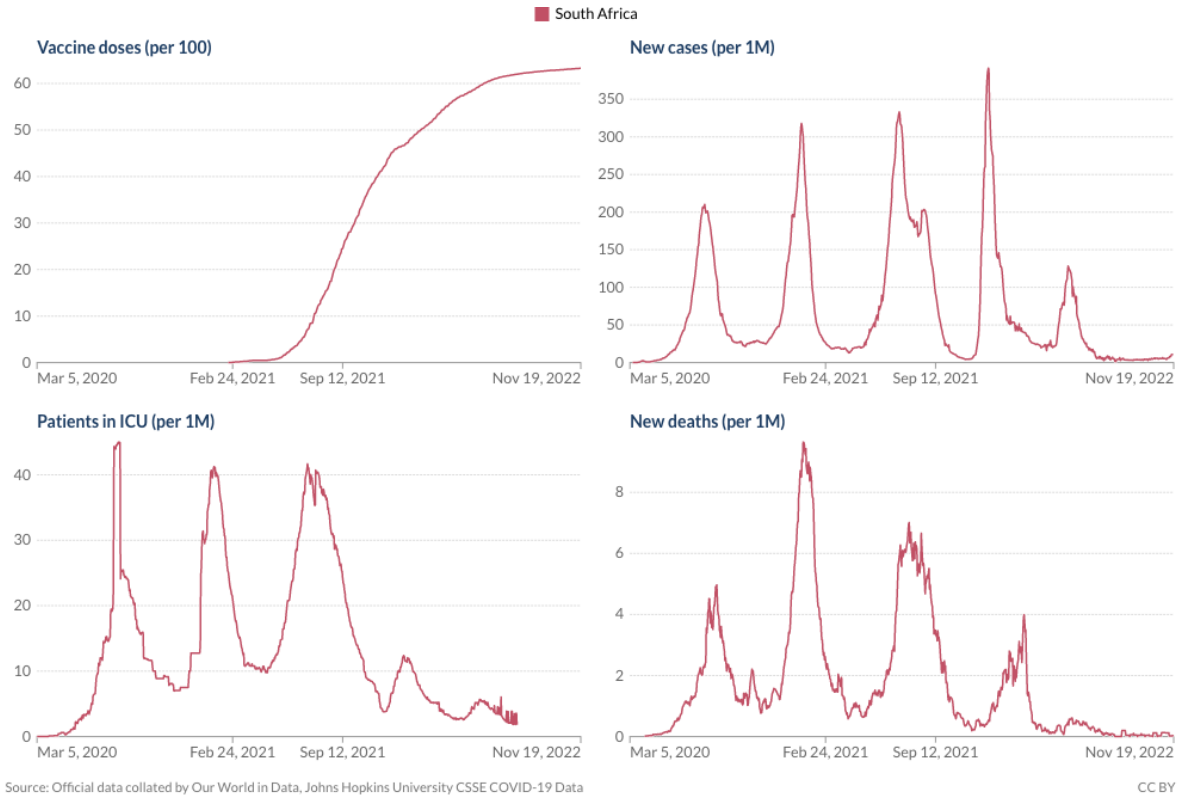






Around July of 2022, the trend lines across both Canada and South Africa began to follow similar patterns in terms of variance diversity as the death rate and infection rate plummeted for both countries. This fell in line with the total number of fully vaccinated individuals

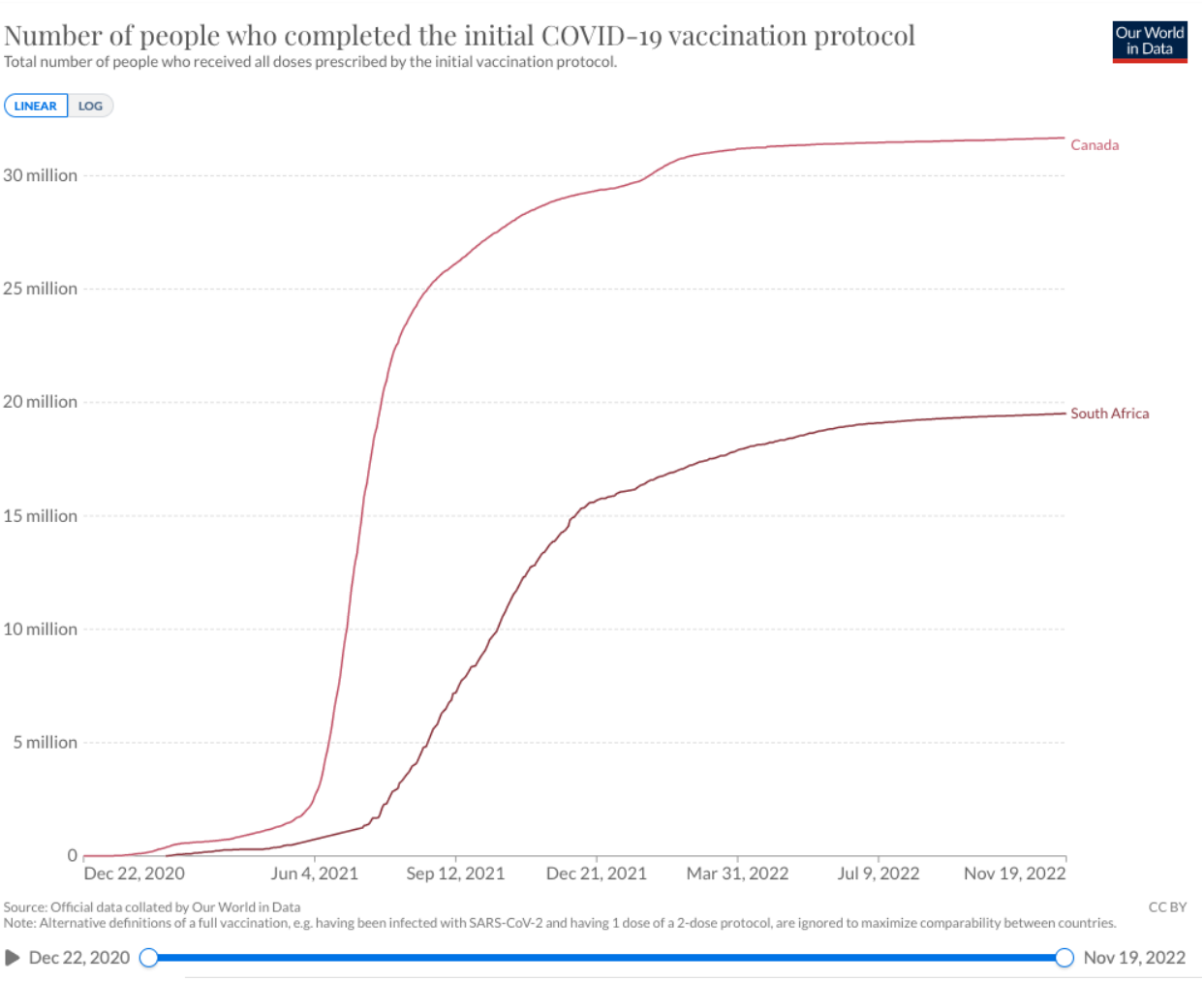
plateauing around a similar time, and the vast majority of induced waves Canada experienced from South Africa's variants came during this delta between their vaccination rates.



While Canadians saw higher vaccine doses, the majority of South Africa's population only received delayed Alpha doses once Delta became the dominant strain, after which there was a spike of vaccinations for the included delta strain in December of 2021.

By that point however, South Africa had more than 96% of its cases come from Omicron while Delta was still the dominant strain in Canada. These mismatches and delays in vaccination policies not only makes countries like South Africa with their disease profile suspect to risk of increased mutability, but further puts the lack of vaccination as a critical policy factor in the spread and development of new variants for countries across the world. As medical reports from Cape Town's hospital networks (including Prince Mshiyeni Hospital) reported, delayed vaccines will have little to no effect on different waves. The vast majority of African nations, including South Africa following the Omicron wave had a small vaccinated population. Fewer than half a million people had received shots in Sub Saharan Africa out of a population of 1.1 billion. These discrepancies highlight a systemic inequality in vaccine availability, where in the US (330 million population), over 90 million were administered doses while more than half of the UK's 67 million population have gotten at least one shot. AstraZeneca PLC's newly issued

vaccine stopped being administered in South Africa as research indicated that the vaccine was ineffective against the current and upcoming growth in new strains. Genetic diversity is thus a central component of the South Africa and Sub-Saharan African battle against COVID-19, and thus vaccinations are the most effective way to prevent highly susceptible populations from facilitating mutations and greater variation. As such, the inclusion of boosters and updated conjugates that countries like Canada and the US have access to makes them more protected against this kind of battle vs. less developed countries like South Africa. While Pfizer and AstraZeneca's vaccines continually prevent hospitalization along with severe illness and death even in new variants (as seen by the drastic decrease in hospitalizations and ICU vistics in South Africa after September 12th despite being subject to a new set of waves), the continual growth rate of cases poses a greater future risk (each new wave has been found to be increasing in total cases but a reduction in deaths while vaccine doses have stagnated, indicating that the vaccine conjugates that target the heavily mutated spike protein are slowly becoming ineffective at stopping the growth rate but do reduce extreme symptoms).

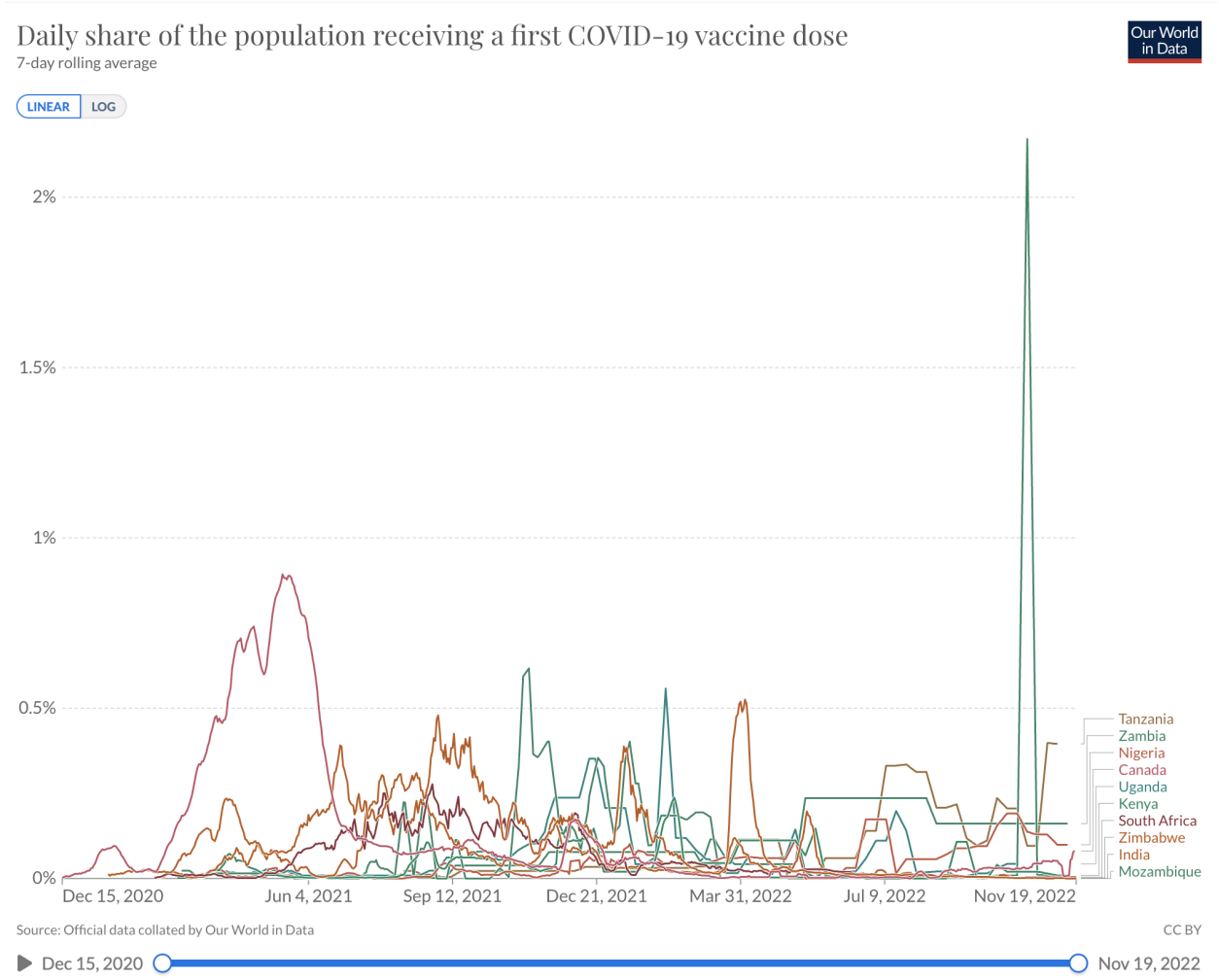


The economic impact of infections is still greatly straining on countries like South Africa, despite having lower deaths.

The mismatch in vaccination patterns also led to viral mutations that made the management of the pandemic more difficult. Both S:E484Q (3237) and S:L452R (5257) mutations were found together as complementary spike RBD neutralization resistant mutations that quickly grew to make up more than 20% of confirmed Omicron positive cases. The substitution mutations had an abnormally-shaped channel protein on the cell membrane that increased viral reactivity and correlated directly with transmissibility.

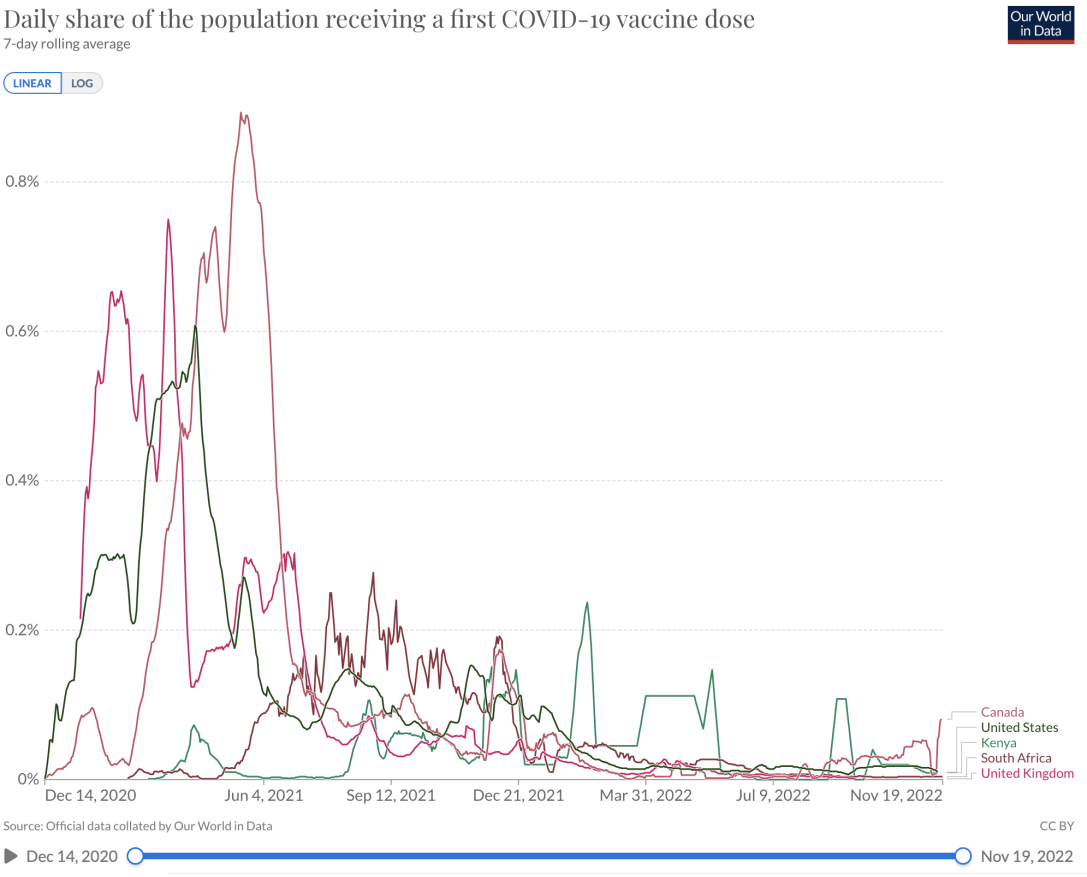
The incidence of this mutation was found to have increased in patients 2 weeks after being administered by the Pfizer vaccine, indicating selective pressure from increased vaccine and monoclonal antibody resistance.

Notice how vaccination rates only spiked mid-2021 onwards for countries like South Africa and Kenya, while many African countries with a high immunocompromised population are yet to have received vaccines. Some of these countries have seen greater spikes prolonging the Alpha-Beta waves, the disparity between vaccination rates and the genetic diversity and dominance of new variants is clear.



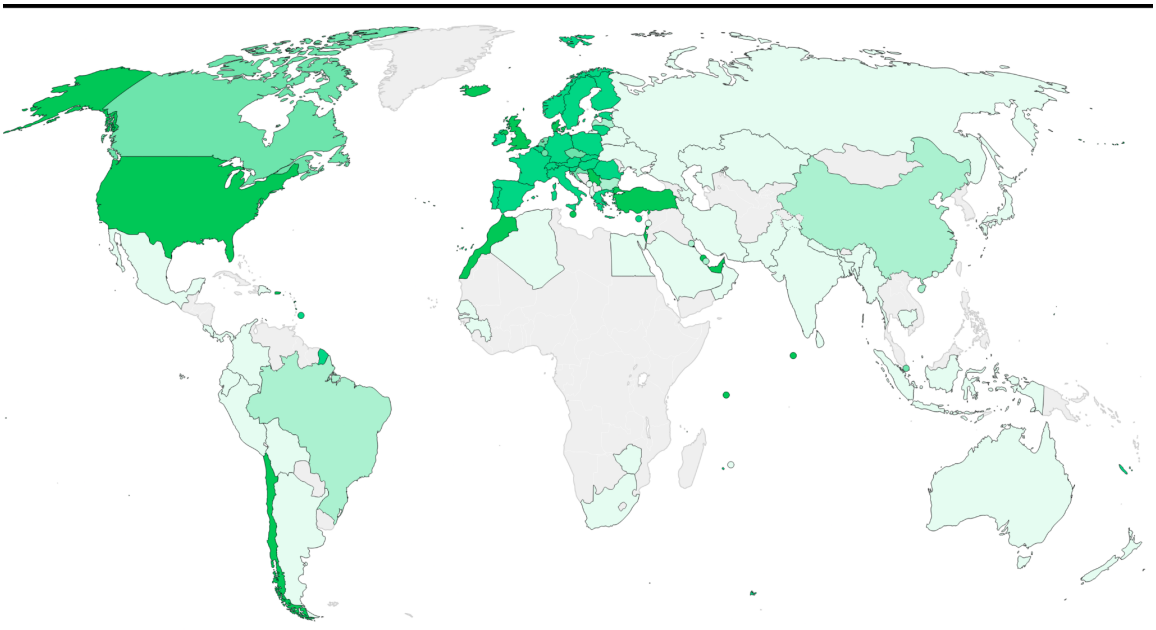
Alternative mAB treatments also failed to address growing Omicron concerns. S:E484A (the most common mutation with 6732 total instances) cited epitope effects found the substitution mutation to show non-selective monoclonal antibody resistance, specifically across 4 unique antibodies with little but substantial resistance against all other antibodies in the assay. This was the first instance of general-immune escape and antibody neutralization behavior in the Omicron lineage. While it is not directly found to be an immune escape mutation, it was highly effective against the commonly cited bamlanivimab mAB therapy, causing it to lose 8x binding affinity. [5]

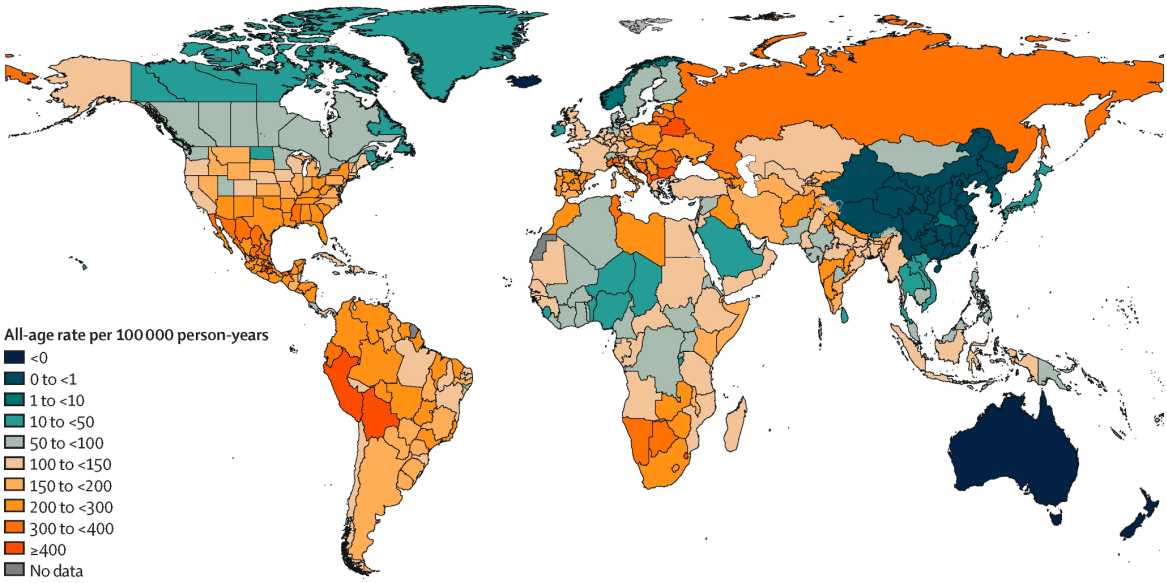
For instance, the vast majority of variants have been seen to originate directly from countries like South Africa, while Northern countries with lower immunocompromised and relatively healthy populations (minus cardiovascular and obesity) get the vast share of their genetic diversity directly from second-hand infections and transmissions. If we look at the genetic makeup of variants and the specific mutation rates of these countries.



A higher vaccination rate could not only reduce the SARS-CoV-2 infection rates, but it can also prevent the severity and length of infections which have been found to correlate with immunocompromised patient infections and the higher rates of mutation, thereby reducing the possibility of new mutations forming. As such, South Africa's vaccine policy should prioritize coverage, speed, and forwardness over treating current waves. This can only be accomplished by direct genomic surveillance and predictive models that can identify possible mutations in the future.

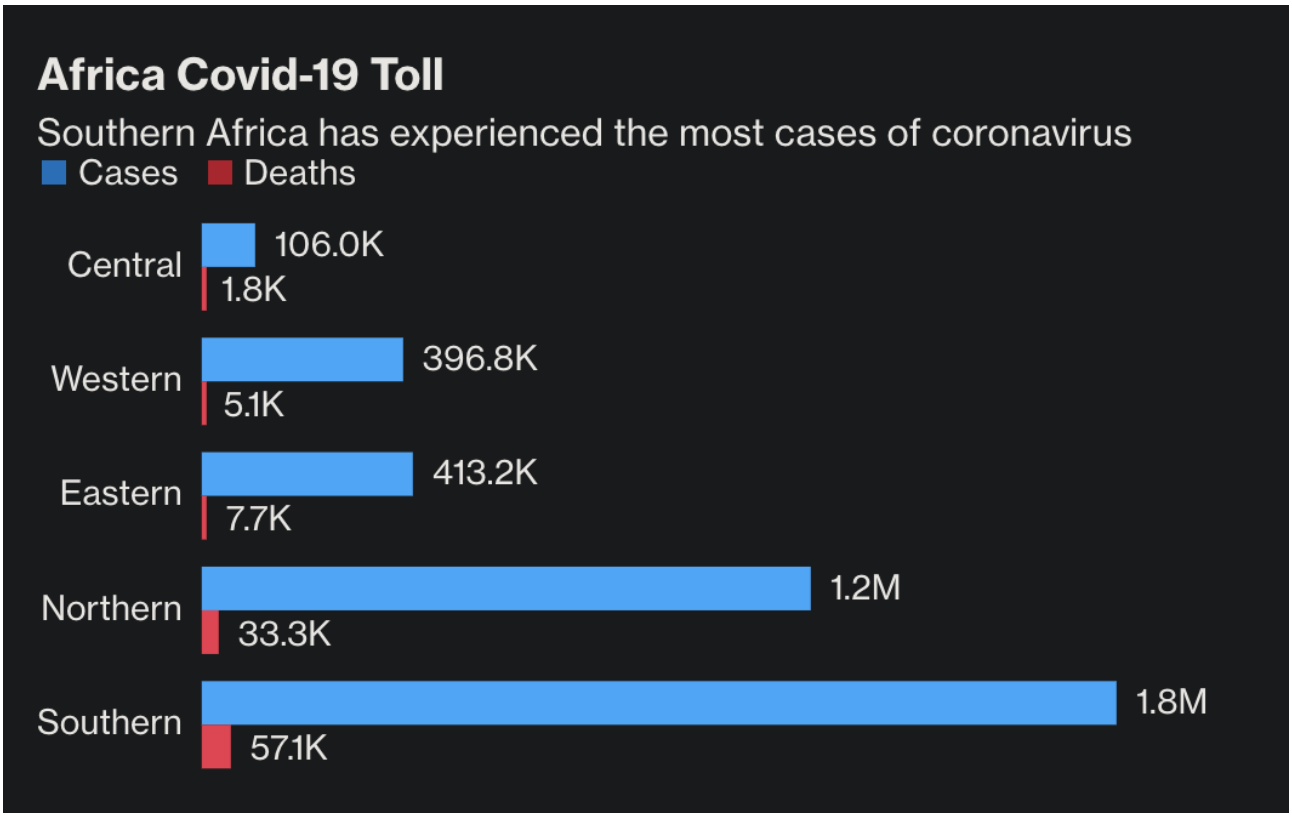
The spread of mutations originating from Sub-Saharan Africa have been pin-pointed as the root cause for new waves and accelerated infections in other countries. Developed nations that have pre-paid for vaccines have the ability to organize faster than most, but vaccinations here do not limit global viral strain diversity and mutability power as the pandemic's epidemiological properties and country's infected demographic pose little to no impact on global pathogen management. More than 75% of total produced vaccines were administered by 10 countries at the start of the pandemic, with SubSaharan Africa seeing the least amount of vaccines. A report by the anti-poverty group stated that South Africa received slightly more vaccines than its neighboring countries because the country has far stronger and more robust medical socioeconomic protections for its citizens with greater healthcare capacity. Nevertheless, the distribution of vaccines does not reflect the immunosuppressed population groups and those at high risk across the world.





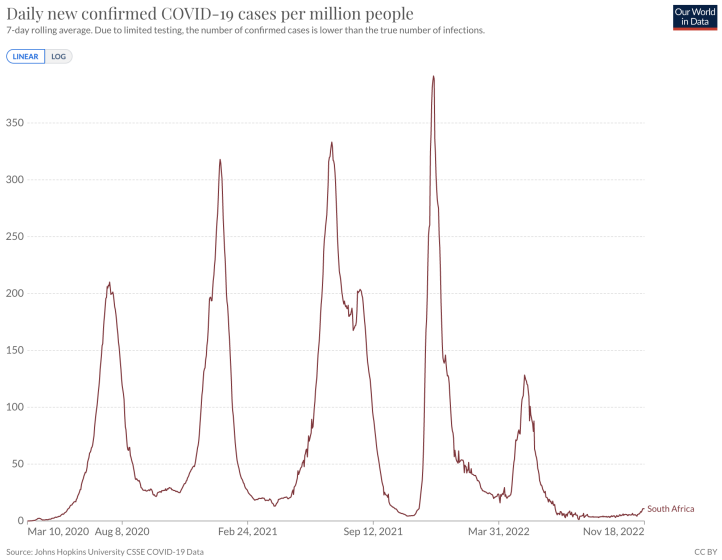
Hotspots of greatest risk (excess mortality) do not correlate with distribution of vaccines, indicating a global mismatch. The world's richest countries seem to be on track to accumulate a billion more doses in excess while Africa's adult population is at risk. The correlation between immunosuppressed populations being highly concentrated in the Southern region with total cases makes this apparent, while South Africa has received some of the least vaccines in the continent.

The primary benefit South Africa has is unlike other nations, it can quickly deploy vaccines through roads, storage systems, and cooling. Doctors have access to oxygen and nurses to administer doses, but a rapid influx would greatly cripple the still vulnerable hospital networks in Cape Town as the total cases increases. In tangent with vaccines, reliable testing data has been limited in the region, and inadequate data has limited the effectiveness of monitoring COVID-19 variants earlier on and then changing the available vaccine conjugates to reflect new coming waves. Genomic surveillance is critical here, but requires tangential vaccine efficacy metrics (particularly for viral vectors and their impact on the rate of infections of corollary viruses like HIV and COVID-19). South Africa continues to provide 11k new daily doses, which would mean that it would take over a decade to cover the 40 million target.

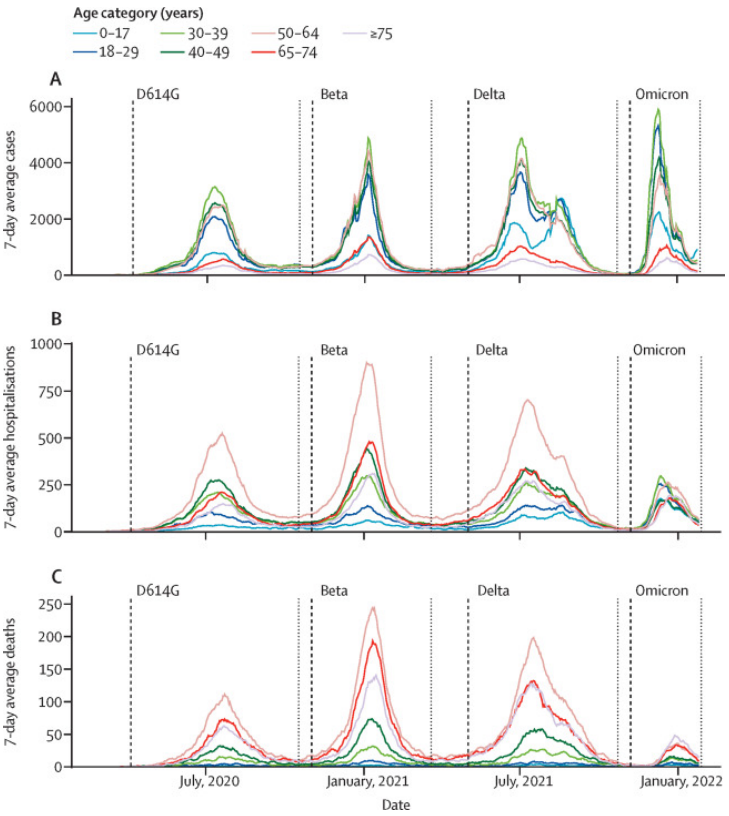


For instance, one of the primary reasons why Omicron cases were reported was because they did not display immediate symptoms, whereas prior waves like Alpha and Delta especially had a greater pathogenic effect. In reference to the objective survival fitness function for SARS-CoV-2, the introduction of vaccines selected for mutations that increased the total time a viral strain can stay in the host cell over immediate pathogenicity. This not only entailed escaping immune responses and reducing sensitivity to T-cell responses as discussed before, but also indicates how evolutionary pressures from external circumstances prompted the specific properties of Omicron. Across Omicron spikes in South Africa, this novel spike protein mutation T478K (most common mutation, found in 10701 cases) responsible for the ACE2 interaction complex was identified in around 86% of B.1.1.222 variants. The doubling rate of this mutation was higher than any other and followed right after the sudden decline of a more popular Beta-variant D6146 mutation that was responsible for increased virulence to promote entry into the host cell. However, the Omicron variant obtained a similar, more superior transmissibility mutation from this substitution of threonine with lysine. This was one of the founding mutations of the Omicron wave that overturned D614G right March 26, 2021 at the start of the Omicron wave. The mutations to the spike protein further allowed the lineage to develop resistance by neutralizing antibodies to evade immune response. Tertiary protein structure analysis revealed that the mutated amino acid resides at the most sensitive region of

the spike protein responsible for interacting with the ACE2 human receptor protein. These mutations are largely influenced by an evolutionary pressure to reduce affinity in sacrifice for longer survival rates, indicating the ability for the virus to stay dormant via immune escape without causing significant symptoms in the host.

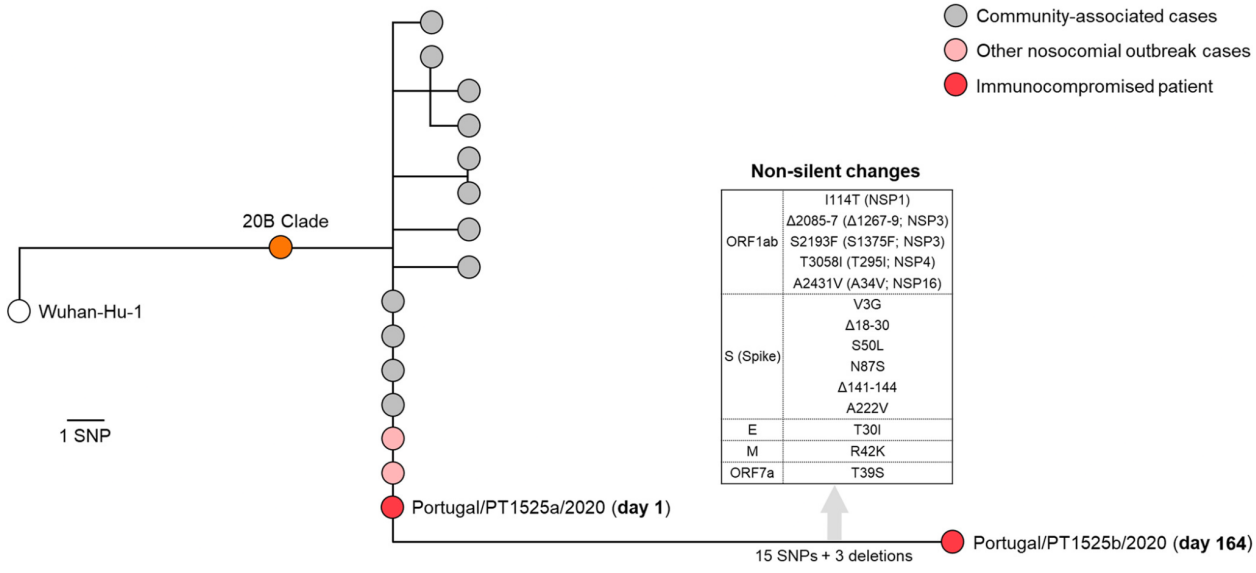


Rank	Country	HIV/AIDS - people living with HIV/AIDS
1	South Africa	7,500,000
2	Mozambique	2,200,000
3	India	2,100,000
4	Nigeria	1,800,000
5	Tanzania	1,700,000
6	Kenya	1,500,000
7	Uganda	1,500,000
8	Zimbabwe	1,400,000
9	Zambia	1,200,000
10	Malawi	1,100,000
11	Russia	1,000,000
12	Brazil	920,000
13	Ethiopia	670,000
14	Indonesia	640,000
15	Congo, Democratic Republic of the	520,000
16	Cameroon	510,000
17	Thailand	480,000
18	Botswana	380,000
19	Lesotho	340,000
20	Ghana	340,000
21	Angola	340,000
22	Ukraine	250,000
23	Burma	240,000
24	Mexico	230,000
25	Rwanda	230,000
26	Vietnam	230,000
27	Namibia	210,000
28	Colombia	200,000
29	Swaziland	200,000
30	France	190,000
31	South Sudan	190,000
32	Pakistan	190,000
33	Romania	190,000
34	Haiti	160,000
35	Spain	150,000
36	Mali	140,000
37	Argentina	140,000
38	Italy	130,000
39	Chad	120,000
40	Togo	120,000



With a greater investment in genomic surveillance, South Africa's established genomic institutions have the potential to monitor immunocompromised individuals and the creation of novel variants in tangent with vaccination system effectiveness with rapid testing. As Tulio de Oliveira (South African Genomics Institute head) stated, "South Africa really risks becoming one of the mutation factories of the world". For instance, epistatic modeling showed that the B.1.1.529 genome variant stood out due to its 30 changes in the spike protein, and was found to heighten ineffectively and ability to evade infection-blocking antibodies. A study that looked at an immunocompromised lymphoma patient found that the SARS-CoV-2 infection marker evolved to have 5 single-nucleotide polymorphisms (SNPs) (11 leading to amino acid alterations) and 3 deletions accumulated during this long-term infection, four amino acid changes (V3G, S50L, N87S, and A222V) and two deletions (18-30del and 141-144del) just at the spike protein.

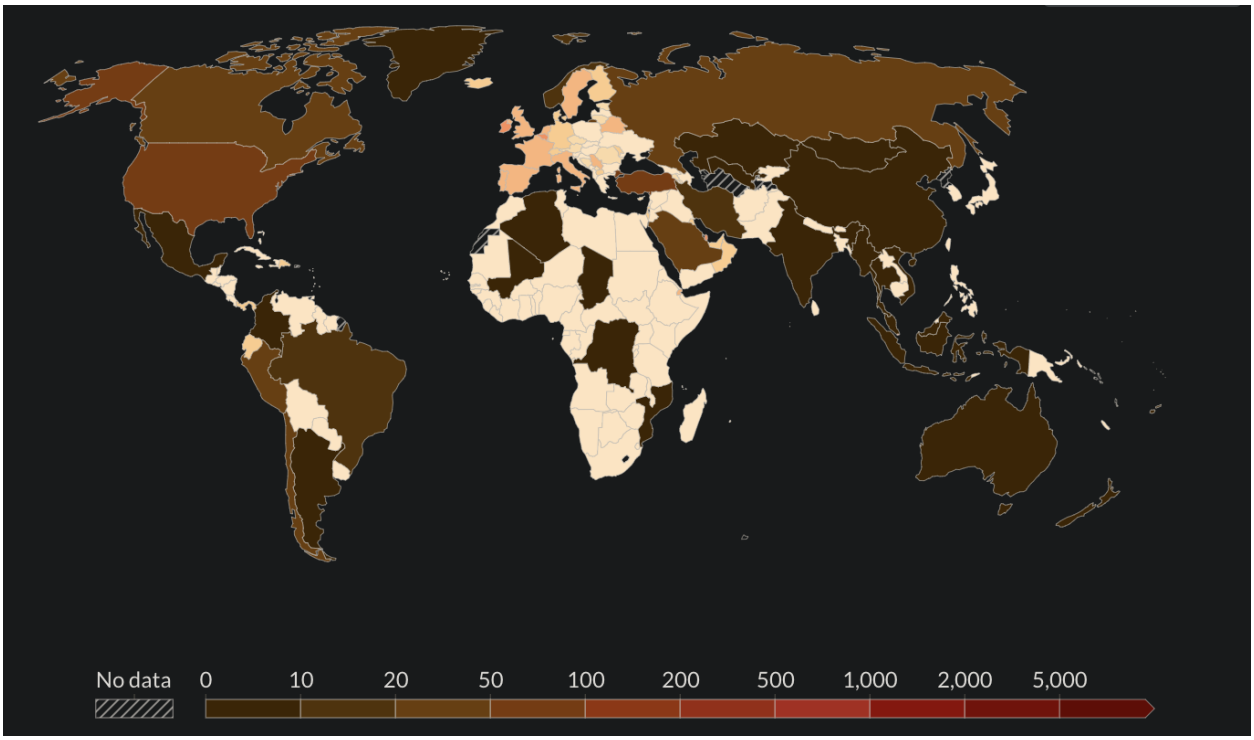
These same mutations were found to have been detected across similarly infected individuals, indicating a kind of convergent evolution amongst different SARS-CoV-2 strains between different hosts. Antiretroviral drugs like remdesivir have been associated with shaping viral population shifts by neutralizing the antibodies in an immunosuppressed patient in response to the same antiretroviral and convalescent plasma treatments. As such, the incidence of HIV not only directly impacts the body's production of antibodies for treating COVID-19, but the treatments and therapies used against HIV establish an evolutionary pressure. Many of these mutations induced from HIV have been found to affect the SARS-CoV-2 affinity to ACE2 receptors, immune evasion, and increase entry efficiency. These abnormally high amino acid change frequencies are directly linked to the presence of HIV, where both viruses saw similar mutations at the transmembrane coat against the body's produced antibodies. Del141-144 emerged in both SARS-CoV-2 and another set of immunocompromised patients with varying diseases (non-Hodgkin lymphoma, antiphospholipid syndrome, and asymptomatic cancer).



The vast majority of these mutations (derived from 77 viral samples) could be detected by genotype tests that result rapidly due to their prevalence on the spike protein. South Africa's current detection policy however is limited to initially symptomatic patients and other key groups with limited opportunities for mass open testing regimes compared to other countries at a similar economic status. Furthermore, South Africa has experienced the most delays and unreliability from the WHO when it comes to rapid antigen testing results. In 2021, the country deployed vaccines first to healthcare workers followed by workers and different age groups before aiming to achieve herd immunity by 2021. HIV not only promotes similar mutations to host cells that increase COVID-19 susceptibility, but also promotes the use of monoclonal antibodies and vaccines. These are staples in South Africa's approach to treating their HIV epidemic amongst all demographics and socioeconomic groups (particularly young women and middle-aged men). These are cost-effective solutions that have passed Phase 2b/3 trials in countries, with many already being approved for HIV prevention. However, mAbs have been directly associated with promoting resistant gene mutations in B.A and B.1 lineages that also give these variants the potential for immune escape. As such, immunocompromised patients and the treatments they are administered have a high overlap with mutations that improve COVID-19 fitness.

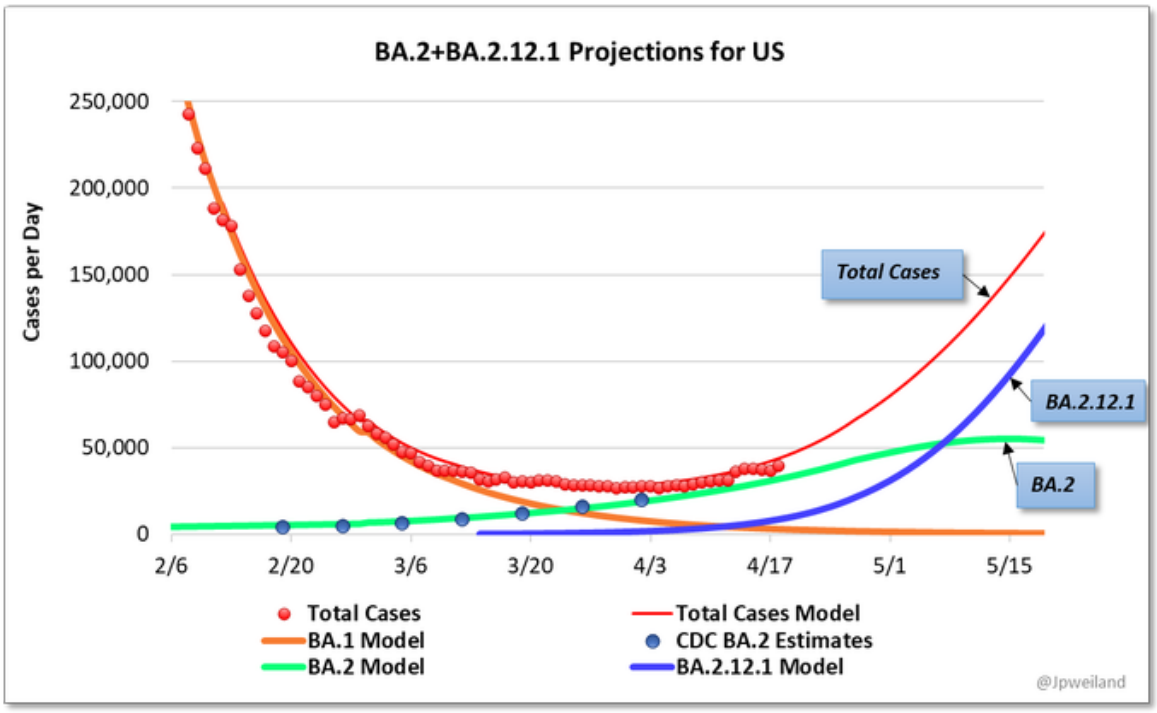
Argument and Observation 2 from Epistatic Modelling: Competition among variants and evolutionary pressures

South Africa's pandemic policies can be assessed in relation to the computational data and methods used to collect insights on causation and correlation of the variant's genetic



mutations, epidemiological properties and dynamics, and serology data. While initial transmissions in South Africa started off low due to their cited lockdown policy from March 26 to April 16. These interventions were quickly relaxed and paired with delayed vaccine rollouts that failed to directly curb the spread of the current VoC with its amplified transmissibility rate, more than 30% of the population would be affected by COVID-19, and then reached up to 63% in 2021. Even still, these lockdown measures were only applied at the beginning of the pandemic when the Alpha variant was more pronounced across other nations and found at a much lesser extent in the greater continent of Africa.

The hallmark traits of the evolved Omicron variants that set off the most recent, largest global COVID-19 wave included low pathogenicity but lower detectability by PCR tests and a 4x transmissibility rate compared to the South-African originating Delta variant. While the specific mechanism by which transmission of viral loads was well understood as the spike protein binds to human ACE2 cell surface receptors in the nasal cavity, there was little to know understanding for the convergent evolutionary pressures that acted as the external stimuli and pressure. However, the genetic analyses done prior in the experiment provided relevant details on the importance of co-existence and competitive nature between VoCs as coupled with the emergent properties of Omicron and its sub-lineages.



Notice how following the presence of the BA.2 clade, the BA.1 model projects a flatlining growth as total cases go down, but BA.2 cases grow exponentially and become almost all of the total cases per day in the United States. This repeated dominance and displacement of variants can be explained by the mutation data found here.

This competitive evolutionary pressure not only sheds light on the nature of the SARS-CoV-2 pandemic occurring in waves, but further supports the argument that viruses are able to quickly manipulate and change their objective fitness functions at a higher rate than most organisms. This is the reason why both Omicron and Delta have almost complementary virological and epidemiological properties from immune erosion to circulation time in the body. Each individual case can be explained using the highest-frequency mutations found in South Africa's viral dynamic data.

Using phylogenetic analyses paired with fusion assays to measure viral infectivity, cleavage efficacy, and fusogenicity, S:S375F was identified to have produced a base-pair substitution that switched the position of the N-terminal domain and Receptor-binding domain of the spike protein, placing it at the end. This was abnormal behavior but seemed to be responsible for the unique virological features of Omicron, particularly the infectivity rate. This is further supported by the fact that total case numbers of the B.1 strains spiked, growing at their highest rate right after the acquisition of this mutation due to the mutation leading to low cleavage efficacy, high population transmission rates, and fusogenicity between ACE2 receptor and the Spike protein. [8]

The BA.2 variant quickly overcame BA.1 with a higher replication fitness and greater pathogenicity, placing itself as the primary VoC when referring to the new Omicron wave. It was found that G496S in the BA.1 spike mutation had a negative effect on its replication fitness, and the lack of this mutation was selected for in BA.2 to allow it to have a lower sensitivity and resistance to therapeutic monoclonal antibodies via immune escape. This outlines the evolutionary pressures at play amongst genetically similar variants.

In a similar fashion, the mutation G446S (1376 cases) explains the shift in Delta VoCs' virological and epidemiological behavior with the suppression of the understood fitness model switch with the introduction of Omicron. This substitution mutation was found at the N-terminus of the spike protein epitope that plays a direct role in interacting with the CD8+ T-cells following vaccine introduction into patients. The mutation prevented the automatic TCR recognition by the immune system in the body and enhanced replication in Omicron. Inhibitor studies were conducted to demonstrate that the G446S mutation directly impacted antigen-presenting markers on the spike protein by T-cell vaccine-invoked immunity. However, what was interesting was that the G446S has an alternative effect on Delta variants where the mutation instead has a positive effect on transporter-associated antigen presentation. As such, T cells were more efficiently able to recognize target cells expressing these mutations in Delta variants even after vaccination. This signaled that the specificity of the spike protein mutation was dependent on secondary mutations based on direct coupling values. The hypothesis proved true, where the presence of an L452R substitution in the Delta spike protein's NF9 peptide increased T-cell sensitivity while this NF9 epitope mutation was not found in Omicron.

This would indicate that the G446S also impacts other cellular processes that suppress T-cell sensitivity while this is not possible in Delta. As such, the addition of this mutation for Omicron also serves as an evolutionary mechanism to increase survivability while other strains like Delta are susceptible to enhanced recognition. Additionally, there is an immunocompromised overlap with this mutation in SARS-CoV-2 and HIV patients. Following antiretroviral therapy, alanine to proline substitutions occurred on HIV residues that triggered epitope changes which were directly responsible for T-cell responses to COVID-19 variants. As such, it is possible immunocompromised patients have accompanying mutations that rapidly increase the rate by which G446S occurs in the spike protein as an evolutionary pressure. For instance, the percentage of suppressed GD137 T-Cells in were higher in those with HIV patients than those without.

In regards to Delta, K417N (4958) was found to be unique to the variant and identified originally in South Africa for its ability to neutralize monoclonal antibody treatments. The doubling rate for the respective Delta strains was also greater than previously mutated strains, indicating higher virulence and transmissibility. The mutation was also commonly found in conjunction with N501Y mutations in the spike protein, where their combination increased RBD-ACE2 interactions at a larger rate and were compatible with other high transmissibility mutations found commonly in South Africa (K417N and L417A).

In terms of transmissibility, there are countless mutations that were found that signified a shift in the parameters SARS-CoV-2 was optimizing for to achieve a higher fitness rate. The S371P substitution mutation (1840 cases) impacts several key metabolic pathways, particularly hemostasis in the body. Particularly, the mutation deregulates receptor signaling pathways and improves signal transduction of S1P Receptor which is responsible for adhesion and transmission. These were commonly found in the BA.1 lineages and was one of the tail-end mutations that came with the increased array of mutations that VoCs encountered. One of the cited evolutionary pressures seemed to be community transmission rates, revealed by docking simulations that the substitution was unique to Omicron and not variants Alpha and Delta. This was found to have been in response to weaker binding affinity between the spike and ACE2 receptors for the 2 earlier variants, indicating competitive fitness pressures.

The beta-core region of the spike protein's receptor binding domain saw its first mutation in the form of a S373P (6732) substitution mutation, and while the mutation was non-synonymous without having changed the secondary structure in any meaningful way, molecular dynamics studies revealed that isomer had more hydrogen bonds present when binding with ACE2 human receptors. As the study reported, this reveals that the S373P mutation offered mechanical stability in the receptor binding domain while simultaneously increasing binding affinity. Viral stability is essential to Omicron variants because slight conformational affinity improvements can dramatically increase the viral load that the body takes in the harsh

conditions of one's airway. The proline substitution particularly worked as well to evade T-cell receptors. The study used protein expression and purification techniques via recombinant restriction enzymes to amplify the samples while analyzing the sequences via direct DNA sanger sequencing. [6]

The most prevalent mutation found to have a direct coupling link between binding affinity and competitive advantage was the spike protein mutation Q498R (6732 cases). Random mutagenic libraries of receptor binding domains were used to identify 1000x fold increases in increased binding affinity with ACE2 from a Q498R spike mutation that was directly epistatic (DCA analysis) with the commonly established N501Y mutation. Similar to T478K and E484k, the mutation is responsible for influencing the receptor-binding motifs and its confirmation when interacting with the ACE2 receptor. Further studies showed that the molecule formed a pi-bond that increased the electrostatic potential of the nucleotide region and created a stronger interaction, leading to increased affinity. Simultaneously, the mutation greatly affected protein stability and binding by 4 fold. This mutation further inhibited competing SARS-CoV-2 variant competition from Alpha, Beta, and Gamma strains. However, this non-synonymous substitution still decreased the clinical disease impact and extreme symptoms of patients, thereby furthering the notion that COVID-19 selects for different kinds of properties by changing its approach to fitness, sacrificing immunogenicity and extremity for long-term dormant survival and also gaining dominance amongst competitive strains. This mutation was often accompanied by S:G496S (3203 cases), and is responsible for the current BA.2 strain's dominance amongst total case numbers.

There are several other ACE2-RBD (receptor binding domain) mutations that were found amongst the most frequent mutation incidences in South Africa's genomic case data. The substitution S477N (6732) was found to have a slight 6% increase in Omicron's ability to bind with the ACE2 receptor, with the frequency of this allele being almost exclusively in African regions. Mutagenesis screens showed some resistance across a broad range of monoclonal antibodies, but particularly for C135 that impacted memory B cell hormonal production, leading to the cell having a decreased binding strength to Omicron variants. In terms of the most recent Omicron strains BA.4 and BA.5, F486V became a significant mutation that allowed novel, upcoming disruptive variants to stay dormant in host bodies before exhibiting pathogenic properties. Deep mutational screening studies revealed that the specific expressed antibodies from the mRNA vaccines were neutralized by the F4867 (1288 cases) mutation, allowing it to give Omicron its high immune erosion % and ability to neutralize current vaccines. While most of the mutations in Omicron BA.1 and BA.2 have enabled it to escape most immune responses, this mutation is especially dangerous because it was found to escape any other remaining antibodies from other closely related variants like BA.4 and BA.5.

Variants of concern become more dominant in total case numbers through their ability to mutate faster than other lineages such that evolutionary pressures can select for the most optimal mutations. Both the spread of variants and antibody recognition via T-cell response effectiveness can cumulatively impact immunity through the mutations outlined above. South Africa is especially of great importance in the greater context of the pandemic as the high immunosuppressed population has enabled certain variants to form by facilitating isolated competition from other lineages. The cases outlined from the genomic analysis data provide a stark contrast to the cross-immunity model that epidemiologists relied on, where competition or any 2 variants closely related on a phylogenetic tree would allow for infected populations to develop immunity for both variants. However, the diverse and rapid rate of mutations experienced in different variants and the wave-like nature that contradicts this co-existing variant model clearly outline certain mutations and their respective phenotypes as being more favorable in out-competing other variants and obtaining a greater share of the total transmissible population. Delta variants typically saw reduced transmissibility because certain pathogenic mutations made their hosts more ill and elicited a greater immune response from the body. The rate at which Delta could mutate to counteract antibody-recognition from the body's T-cells was not quick enough (Delta had a much lower doubling rate and % infection rate while having the highest % immune erosion rate), ultimately leading to the selection of Omicron-variants that favored longer survivability to control a greater share of the total transmissible population. This can be attributed to the Alpha, Beta, Delta, and Gamma variants all having similar pathogenic properties and induced symptoms which promoted cross-immunity dynamics. As such, Omicron and subsequent variants saw less mutations in the more conserved epitopes surrounding immunogenicity and extremity of the infection because it served little evolutionary benefit given that the vast majority of possible desensitizing mutations were unable to counteract the repeated boosters and vaccine efforts, paired with the fact that the possibility of mutations in this space were limited to none. As such, Omicron mutations regressed to counteract the increased immune response by increasing the total viral load exercised by increasing RBD-ACE2 efficacy while staying in the host for longer periods for longer. [12] These properties help explain why Omicron has an incredibly high total case count but little to no deaths. Furthermore, the vaccine antibodies provide an evolutionary pressure for selecting more highly mutable regions in the DNA, which was seen clearly by the overlap in higher mutation frequencies in the total population being at high mutability sites in the amino acid sequence while having a less pronounced DCA epistatic effect. This would allow for subtle but frequent mutations to occur in less conserved regions like the spike protein's N-terminus in Omicron compared to the more common mutations that were noticed in extreme, lesser transmission variants like Delta, to promote the possibility of reinfection in light of new vaccines.

Conclusion

This investigation has a significant role in providing recommendations that can bridge genetic and epidemiological findings with public health policy, especially in vulnerable regions. The primary geographic focus is South Africa as their initial response to the Omicron variant serves as an important case study when it comes to analyzing how the country approached rising cases but little to no induced deaths. A comparison of these factors across different countries in Africa and the rest of the world will provide a useful benchmark in evaluating how these mutations differ in severity and impact. As such, the goal of this inquiry will be to establish novel connections between variants marked by SARS-CoV2 spike protein mutations in different countries and their epidemiological impact.

Sources and Research