Pandemic preparedness through genomic surveillance: Overview of mutations in SARS-CoV-2 over the course of COVID-19 outbreak

Fares Z. Najar,[1,†] *Chelsea L. Murphy,*[1,†] *Evan Linde,*[1] *Veniamin A. Borin,*[2] *Huan Wang,*[3] *Shozeb Haider,*[3,4] *Pratul K. Agarwal*[1,2,*]

[1]High-Performance Computing Center, Oklahoma State University, Stillwater, Oklahoma, [2]Department of Physiological Sciences, Oklahoma State University, Stillwater, Oklahoma, [3]University College London School of Pharmacy, Pharmaceutical and Biological Chemistry, London, United Kingdom, [4]University College London Centre for Advanced Research Computing, London, United Kingdom

†These authors contributed equally to this work.

*Corresponding author: pratul.agarwal@okstate.edu, 405 744-6639

MS 106, Oklahoma State University, Stillwater OK 74078

ABSTRACT

Genomic surveillance is a vital strategy for preparedness against the spread of infectious diseases and to aid in development of new treatments. In an unprecedented effort, millions of samples from COVID-19 patients have been sequenced worldwide for SARS-CoV-2. Using more than 8 million sequences that are currently available in GenBank's SARS-CoV-2 database, we report a comprehensive overview of mutations in all 26 proteins and open reading frames (ORFs) from the virus. The results indicate that the spike protein, NSP6, nucleocapsid protein, envelope protein and ORF7b have shown the highest mutational propensities so far (in that order). In particular, the spike protein has shown rapid acceleration in mutations in the post-vaccination

period. Monitoring the rate of non-synonymous mutations (K_a) provides a fairly reliable signal for genomic surveillance, successfully predicting surges in 2022. Further, the external proteins (spike, membrane, envelope, and nucleocapsid proteins) show a significant number of mutations compared to the NSPs. Interestingly, these four proteins showed significant changes in K_a typically 2 to 4 weeks before the increase in number of human infections ("surges"). Therefore, our analysis provides real time surveillance of mutations of SARS-CoV-2, accessible through the project website http://pandemics.okstate.edu/covid19/. Based on ongoing mutation trends of the virus, predictions of what proteins are likely to mutate next are also made possible by our approach. The proposed framework is general and is thus applicable to other pathogens. The approach is fully automated and provides the needed genomic surveillance to address a fast-moving pandemic such as COVID-19.

INTRODUCTION

COVID-19, caused by the SARS-CoV-2 virus, ground the entire world to a halt in a matter of a few weeks. It aptly revealed that an airborne virus with the ability to quickly mutate, combined with high levels of transmission and fatality, can rapidly cause a world-wide pandemic. The response from the medical and research community was swift, leading to the development of mRNA-based vaccines and several antiviral drugs. Unfortunately, even after three years of this worldwide pandemic, the situation remains alarming. New variants have emerged and continue to emerge, causing new waves of infections and fatality. Our ability to respond to the future pandemics, will depend on our ability to stay ahead of the pathogens. Therefore, there is wide interest in approaches for genomic surveillance as a means for pandemic preparedness and response. To this end, several million SARS-CoV-2 sequences have already been obtained in a hitherto unprecedented global effort. These sequences are being deposited by various health labs in public repositories such as the NCBI's GenBank and GISAID.

SARS-CoV-2 was first sequenced in 2020 from a sample in Wuhan, China. This sequence, widely known as the "Wuhan" sequence,[1] showed that SARS-CoV-2 is a positive sense RNA virus that encodes 26 proteins (Figure 1). The expressed proteins include: four structural proteins known as the envelope, the membrane protein, nucleocapsid protein, and the spike protein;[2] 16 non-structural proteins (NSP 1–16); and six accessory proteins as open reading frames (ORFs), 3a, 6, 7a, 7b, 8, and 10.[3] The two large overlapping open read frames, ORF1a and ORF1b, are transcribed and later split into different NSPs. The overlapping occurs at a frameshift to allow the continuous transcription of NSP12.[1]

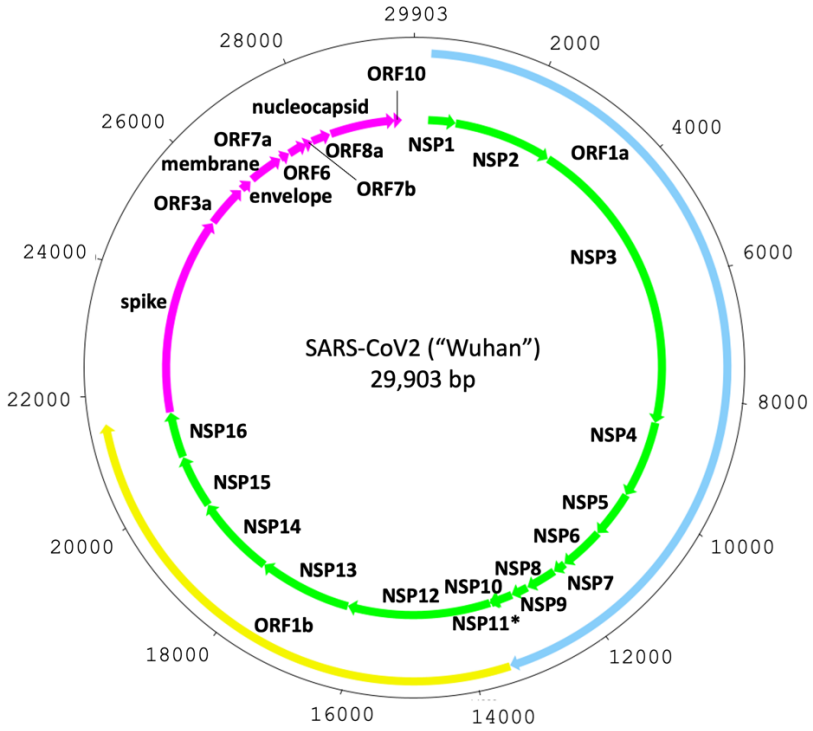

Figure 1: A schematic overview of the SARS-CoV-2 genome. Information based on the "Wuhan" reference sequence. The non-structural genes are shown in green while the structural genes are marked in magenta. The ORF1a and ORF1b boundaries are shown in light blue and yellow respectively. The position of NSP11 (marked by *), which is too small to be visible at this schematic's scale, is not shown.

For the genomic surveillance to be useful for pandemic (such as COVID-19) preparedness, several aspects are required. First, the SARS-CoV-2 proteins that are rapidly changing need to be identified in real-time (or near real-time). Ideally, this information also needs to be available for different geographic locations separately, as the outbreaks show different regional trends. Information on how different variants are mutating is also important, as at any given time there can be (and have been) different variants of the virus in the host population. Collectively, this set of information should enable prediction of what mutated versions of the protein will be presented in the next variants of concern (VOC). However, for this

information to be useful beyond the research community, such as aiding the medical community in making practical but critical decisions, the genomic surveillance should be able to clearly predict future outbreaks or increase in number of cases of infections ("surges"). Furthermore, in order to make an impact on mitigation strategies, it is vital that all of this analysis needs to be performed in real-time.

From a molecular perspective, a detailed analysis of mutations in the virus proteins are needed for understanding of the new outbreaks. Number and types of mutations in individual proteins associated with various variants could shed light on severity of the ongoing or future outbreaks. This information could also help in designing new vaccines and drugs, and combat antiviral resistance which could arise from new mutations. Furthermore, the effectiveness of the antibody-based viral detection methods is also prone to large mutational changes in the different proteins. (Note that some of this information has been generated in-house by vaccine and drug developing groups; however, access to such information remains lacking for the broader scientific community at large.) A detailed understanding of mutational propensities of various proteins would be the key to pre-emptively designing antibodies that will be able to detect mutated proteins, as well as vaccines that are effective against new VOCs. Similarly, relating mutations, molecular structure, and functional mechanism could also provide new opportunities for effective antiviral drugs.

Recently, we described our real-time genomic surveillance of SARS-CoV-2 based on mutation analysis of viral proteins as a methodology for *a priori* determination of surge in number of infection cases. The results are available at http://pandemics.okstate.edu/covid19/, and are updated daily as new virus sequences become available. At present, over 8 million SARS-CoV-2 genome sequences collected all over the world are available from GenBank

(https://www.ncbi.nlm.nih.gov/sars-cov-2/) and were used for analysis of 26 SARS-CoV-2 genes. As a result of this effort we were able to predict the upcoming Omicron BA.5 related surge in end of June 2022, which was confirmed in the upcoming weeks of July 2022. We also issued a warning in September 2022, which was confirmed by infections increase in Europe and several individual European countries. Furthermore, a new watch was issued in the first week of January 2023, which corresponded to the current late January 2023 surge.

Here, we describe the detailed mutations overview of all the 26 proteins and ORFs of SARS-CoV-2 over the course of the COVID-19 outbreak. Our analysis is based on the unprecedented SARS-CoV-2 sequencing campaign undertaken by the medical community and public health labs worldwide. Furthermore, the information available for the number of daily infection cases from public databases has allowed us to make real-time prediction about impending surges in infections. Our approach is model-free, which allows for easy interpretation by the research community as well as the medical community and public health agencies for preparedness purposes. Based on the results, we have also developed an approach to predicting which heavily mutated proteins will likely be present in the upcoming variants. The presented approach and framework is general and can be applied to other future infectious diseases.

METHODS

The raw SARS-CoV-2 genomic sequences data and the number of COVID-19 infection cases are continually obtained from the sources described below. The genomic sequences are first carefully filtered for quality control, and sequences passing quality control are used for identification of mutations and calculations of non-synonymous and synonymous mutation rates and their ratio for each of the listed 26 proteins separately.

Data and data sources: Data for the number of reported COVID-19 cases is accessed from Our World In Data project (https://ourworldindata.org/coronavirus-source-data).[4]

Genomic sequence data: An in-house pipeline of scripts (using Linux commands) was designed around the eUtils tools[5] from NCBI in order to download and process the SARS-CoV-2 records from NCBI's GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Briefly, we used *esearch* and *efetch* commands to obtain these GenBank records. The search string "SARS-CoV-2", refined to "SARS-CoV-2 [ORGN]", was used to download the identified records in the GenBank text format. After workflow optimization post May 2022, the search process used NCBI's newer *datasets* and *dataformat* command-line tools to identify sequences of interest, while continuing to use the *efetch* tool to download records in the GenBank text format. Collectively, a total of 8,026,343 records were searched, and a total of 3,596,381 sequences matching the search criterion were downloaded and used as of July 20th, 2023.

Quality control: Incomplete and ambiguous SARS-CoV-2 genomic sequences and records containing incomplete collection dates were filtered out using the designed pipeline. For the records passing the quality control steps, the nucleotide sequence for each gene was extracted. A non-redundant set of the extracted nucleotide sequences was derived and translated to the cognate amino acid sequences. In the final phase of the pipeline, the accession numbers for each viral isolate, the nucleotide sequences, the associated protein sequences, the collection dates, and the country of collection were stored in an internal SQLite relational database, where they were indexed with unique identifiers to allow for the retrieval and analysis of any part of the parsed data.

Frequency of data updates: As of April 2022, the described sources are being monitored daily for updates. New data is continually downloaded and used for analysis through automated pipelines.

Amino acid alignments and mutational propensity: For each protein, we used Clustal-Omega[6] to align the amino acids from each variant. The total number amino acid changes (substitutions and deletions) were calculated and summed up for each gene. Those values were divided by the total amino acid length for each gene to obtain the mutational rank, where higher mutational ranks correspond to a higher propensity for mutations for that gene.

Alignments and non-synonymous (K_a), synonymous (K_s) and K_a/K_s ratio calculations: The translated protein and nucleotide sequences were aligned using *clustal-omega*[6] and *Pal2Nal*[7] programs to align the codons with their associated amino acids. The resulting alignments were then processed through the program *kaks_calculator*[8] to calculate non-synonymous (K_a) and synonymous (K_s) values and their ratio K_a/K_s values which were used to assess the mutational adaptation for each protein. The parameters required for the *kaks_calculator* program were based on the maximum-likelihood method derived from the work of Goldman and Yang.[9] The first reported SARS-CoV-2 genomic sequence (the "Wuhan" sequence) was used as a reference for all the K_a, K_s and K_a/K_s calculations. While we explored the possibility of using other sequence(s) as references (e.g., the previous day or the previous month), the increasing number of variations available each day makes it prohibitive to select a representative consensus sequence on an ongoing basis. Additionally, we found that using the Wuhan sequence as a reference provided the most intuitive and interpretable results. We observed the mean K_a, K_s, and K_a/K_s over the entire documented timeline of COVID-19 cases to discover if there is a

correlation between these parameters' measurements and the infection surges, and to help reveal the genotypic and virulence natures of the variants of interest preceding the surges.

The targeting of both the non-synonymous and synonymous mutation values was selected due to its ability to observe changes at both the protein and the nucleotide levels. For example, an increase in K_a values suggests a change in some of the amino acids, and can be observed at the protein sequence level, while an increase in the K_s value can only be seen at the nucleotide sequence level. The K_a/K_s ratio, very generally speaking, indicates the nature of the gene adaptation in reference to its predecessor; in this case, the Wuhan genes. If this ratio is greater than 1, it indicates that the gene is going through adaptational changes, or positive selection. This might imply that the gene is responding and adapting to environmental factors. On the other hand, a ratio less than 1 indicates a purifying, or stabilizing selection. In other words, it is *resisting* mutational changes, possibly indicating dependence on the amino acid sequence for function. A ratio equal to 1 indicates no change, or neutral selection. It is important to note that these interpretations are overly simplistic and are provided for the general reader who is not familiar with the concept; in reality, the situation is influenced by many factors, including biases. For example, some of those biases could be transition/transversion bias and codon-usage bias.[10] However, for the purposes of this study, we investigated the K_a, K_s, and K_a/K_s in genes in SARS-CoV-2 sequences collected over the course of COVID-19 as a method for identifying mutational behavior and predicting future surges.

Daily mean values: Each of the tracked quantities (K_a, K_s, and K_a/K_s) have a number of different values associated with them calculated from the different unique sequences reported on a given day. For the ease of tracking the behavior over time, a weighted mean value was calculated for each day. Each of the accession records was assigned to a bin with a unique

sequence. The calculated values were weighted by the number of sequences (records) in each of these bins. Daily average values of K_a and K_s were calculated directly. However, two methods were used for the K_a/K_s ratio.

Two alternate approaches for daily K_a/K_s mean values:

1. A direct approach was first used to calculate K_a/K_s value for each unique sequence if the K_s value was not equal to 0 (the ratio K_a/K_s is undefined when K_s is 0). However, it was found that in many cases near-zero values of K_s led to artificially inflated K_a/K_s values that failed to intuitively communicate mutation trends. The daily weighted mean values of K_a/K_s were abnormally weighted by these sequences with very small K_s values, resulting in very noisy behavior as a function of time. Furthermore, the ongoing curves corresponding to the K_a/K_s ratio did not correspond to mean K_a and mean K_s values. It should be noted that this is not a discrepancy, but rather a consequence of the mathematically necessary exclusion of the sequences with a K_s value of 0. The results from this approach are in the supporting information and the values are denoted as K_a/K_s'.
2. As an alternative approach, we computed a daily value of K_a/K_s by dividing the daily weighted mean value of K_a by the daily weighted mean value of K_s. These values provide more easily interpretable results and are shown in the main manuscript.

Forecasting future surges: The genomic surveillance data, in combination with the number of infection cases, was used to develop a monitoring approach and *a priori* forecasting of an increase in number of infections (or "surges"). K_a, K_s, and K_a/K_s profiles were compared to the infection profile. As we previously described, the ratio of non-synonymous to synonymous mutations (K_a/K_s), which is typically used, did not provide clear trends. Furthermore, the rate of

mutations (a derivative of observed mutations with respect to time) also did not provide a reliable prediction signal. However, we found that collective non-synonymous mutations in key proteins of SARS-CoV-2 showed a significant increase 2-4 weeks before the rapid rise in COVID-19 cases, particularly related to the surges that occurred after the emergence of the Gamma, Delta, Omicron, and BA.5 variants.

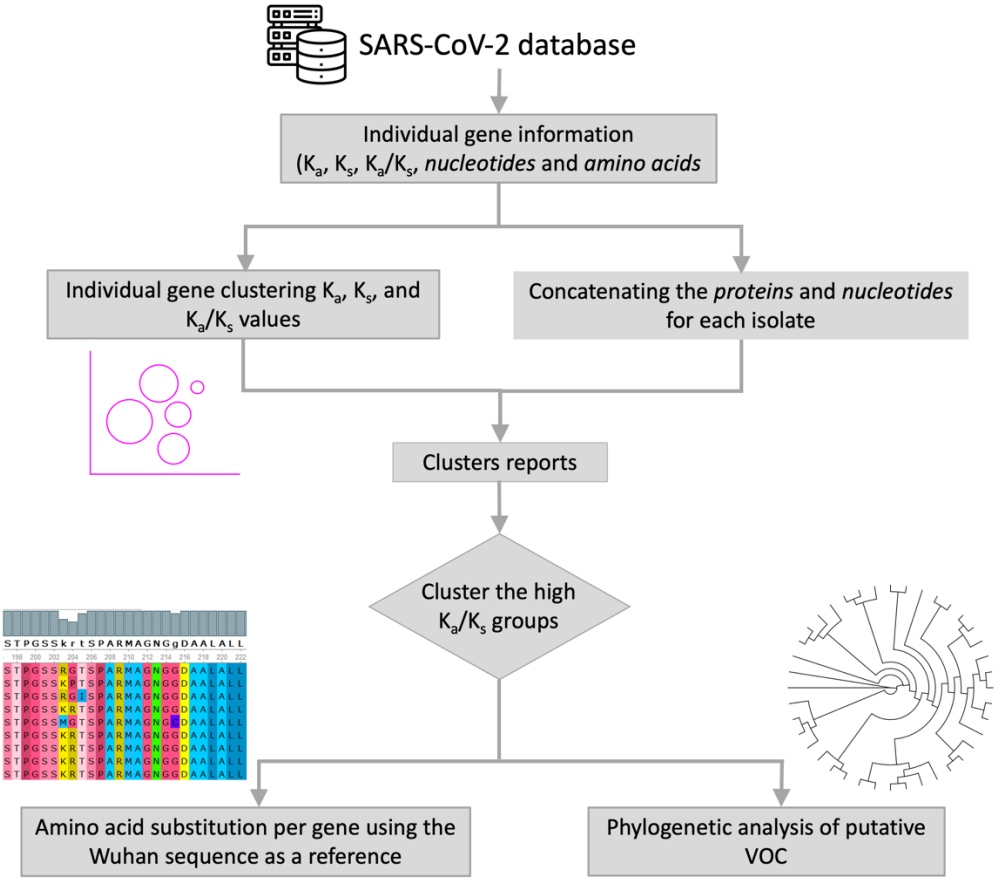

Figure 2: A schematic overview of the approach for predicting future mutations.

Forecasting future make-up of variants of concern: A method to monitor the potential variants of concern (VOC) is described in Figure 2. It is based on examining isolates with genes that have K_a/K_s values larger than one over a sliding 14-day window. First, the gene information for all available variants for the past 14 days are extracted, including the nucleotide and amino

acid sequences and the calculated K_a, K_s, and their ratio values. Next, we group the data for each gene using groups of identical K_a, K_s, and K_a/K_s values. The largest group for each gene is selected for clustering using mmseq2.[11] Non-redundant sequences are then used for phylogenetic analysis. The potential variants are then aligned with the previous VOCs, including the reference Wuhan strain, and the amino acid substitutions are deduced using clustal-omega.[6]

RESULTS

Mutational propensity of various genes: An overview of the substitutions at the nucleotide and protein level of the 26 SARS-CoV-2 proteins, since the beginning of the pandemic are depicted in Figure 3. These results are based on 8.03 million sequences sourced from GenBank ranging from the earliest days of COVID-19 to the present day. The number of quality-controlled accession records range from 2.2 million to 3.5 million for different genes (see Table S1 in the supporting information for details for each gene). The number of unique sequences as a proportion of total daily sequences ranges from almost 0% to 10.6% at the nucleotide level (Figure 3A), while at the protein level it ranges from 0.0% to 5.0% (Figure 3B). Note that the genes are listed in decreasing order of unique sequences. As expected, some of the higher numbers of unique sequences are observed for the large genes and proteins while smaller genes show some of the lowest numbers of unique sequences; however, the trends are in the middle are not uniform. Additionally, the percentages of unique proteins versus nucleotide sequences for each gene differ vastly. There is a higher percentage of unique nucleotide sequences as compared to their amino acid counterparts, indicating that most of the substitutions are synonymous.

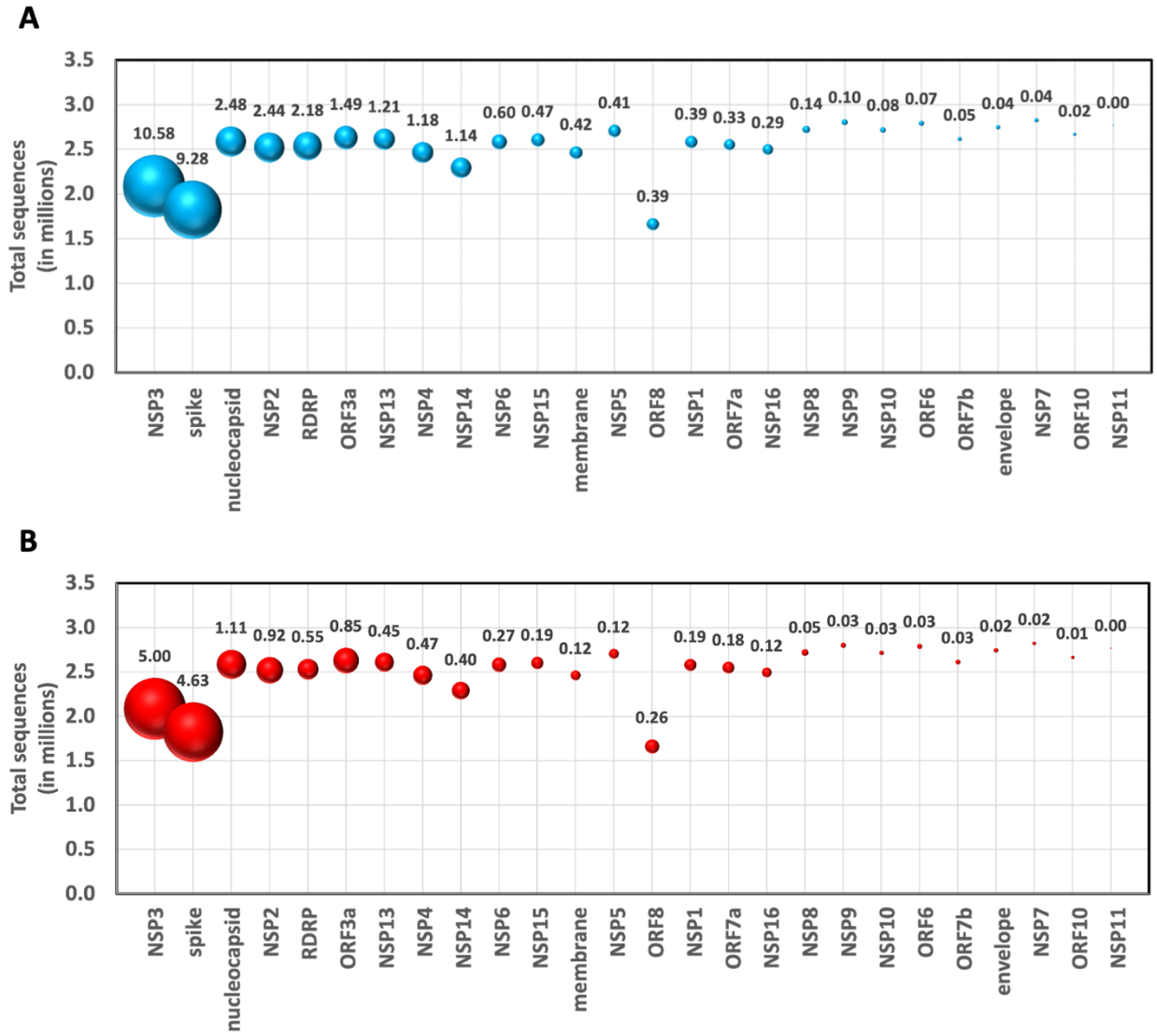

Figure 3: The percent of unique sequences for each gene. (A) Unique nucleotide sequences. (B) Unique protein sequences. The size of the sphere corresponds to the relative percentage of unique sequences for each gene and the percentages of unique sequences are labeled.

Mutational propensity, defined as the total number of unique mutations for each gene divided by the total length (in amino acids), displays a wide range across the genes and provides several insights (Figure 4, depicted as percentage). In the collected sequences, the spike protein, NSP6, and the nucleocapsid protein have shown the largest propensity to mutate. The spike protein is associated with viral membrane structure, while NSP6 is associated with the host's

endoplasmic reticulum membrane. Additionally, the four external proteins (spike, nucleocapsid, envelope, and membrane proteins) rank in the top six highest propensities. There are several other notable observations; most of the proteins with the lowest ranks (indicating proteins that have been more mutationally stable so far) seem to form complexes with other proteins. For example, NSP7 and NSP8 form a complex with NSP12 as part of the replication complex. Similarly, NSP10 and NSP14 are needed to activate NSP16, with NSP10 also serving as an activator of NSP14. Another notable observation is that the top-ranking proteins in mutational propensity are membrane-associated protein. This analysis can potentially be useful for protein targeting for vaccine and for drug development.

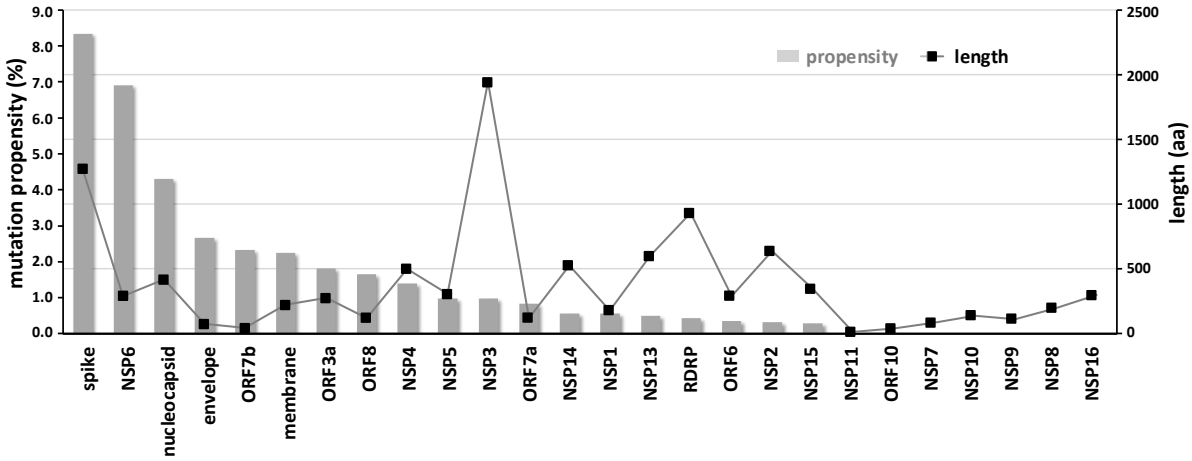

Figure 4: Mutational propensity of various SARS-CoV-2 genes. The number of mutations observed in each gene at the protein sequence level and normalized by sequence length is depicted (as gray bars). The genes are ranked in order of their mutational propensity. The length of each gene, shown as black lines, shows no correlation with the mutational propensity. See the table in the supporting information for the exact values.

In the remaining section, we discuss the important observations for each individual protein. We start with spike protein, as it has shown the most interesting behavior. Then we describe the three other structural proteins, followed by NSPs and the various ORFs.

Spike Protein. This protein plays a vital role in the successful internalization of the virus into the host cells by interacting with the host cell angiotensin-converting enzyme 2 (ACE2) receptor.[12] The protein consists of three domains: a receptor-binding domain that binds to the human ACE2, a fusion domain, and a transmembrane domain.[13] Cleavage of the spike protein by host proteases kickstarts the process of uptake into the host cell.[14] The spike protein has been the predominant target of mRNA vaccine development. Recently, it has been hypothesized that the spike protein may interfere with the immune synapse assembly and function that is normally responsible for coordinating the killing of virally-infected cells by cytotoxic T lymphocytes.[15]

Around 50 different mutations have been observed in the spike protein since the start of the pandemic (see Table 1 and Figure 5). Note that for our analysis, we used only the fully quality controlled sequences. As described in Table S1 in the supporting information, there are almost 2.36 million records, with a total of 227,602 unique nucleotide sequences and 115,658 unique protein sequences for this protein. Different mutations are present in all VOC but the number of mutations significantly increased beginning with the Omicron BA.1 variant in the post-vaccination period. As visible from Table 1, compared with all the other genes, the spike protein has the greatest number of observed mutations in all the SARS-CoV-2 variants. Mutations marked with an asterisk (*) are of particular interest, as these amino acids interact with ACE2 receptor, which are present in most VOCs, and again have increased since the emergence of Omicron BA.1 variant. Functionally, these mutations are expected to affect the interaction with the ACE2 receptor as a number of them are located in the receptor binding domain (Figure 5).

Table 1: Summary of all mutations observed in the 26 SARS-CoV-2 proteins

Gene	Alpha	Beta	Gamma	Delta	Omicron BA.1	Omicron BA.2	Omicron BA.3	Omicron BA.4	Omicron BA.5	Omicron XBB.1	Omicron XBB.2	Omicron XBB.3	Omicron XBB.4
Spike	Δ69[a,b]	D80A	L18F	T19R	A67V	T19I	T19I	V3G	T19I	T19I	T19I	T19I	T19I
	Δ70	D215G	T20N	L452R	Δ69	Δ24	Δ24	T19I	Δ24	Δ24	Δ24	Δ24	Δ24
	Δ145	Δ241	P26S	T478K	Δ70	Δ25	Δ25	Δ24	Δ25	Δ25	Δ25	Δ25	Δ25
	N501Y*	Δ242	D138Y	D614G	T95I	Δ26	Δ26	Δ25	Δ26	Δ26	Δ26	Δ26	Δ26
	A570D	Δ243	R190S	D950N	G142D	A27S	A27S	Δ26	A27S	A27S	A27S	A27S	A27S
	D614G	A264S	K417T	A1174V	Δ143	G142D	S477N*	A27S	Δ69	V83A	V83A	V83A	V83A
	P681H	K417N	E484K*		Δ144	V213G	T478K	Δ69	Δ70	G142D	G142D	G142D	G142D
	T716I	E484K*	N501Y*		Δ145	D405N	E484A*	Δ70	T478K	D@145	D@145	D@145	D@145
	S982A	N501Y*	D614G		N211I	R408S	Q493R*	G142D	E484A*	H146Q	H146K	H146K	H146Q
	D1118H	D614G	H655Y		L212V	K417N	Q498R*	V213G	F486V*	Q183E	Q183E	Q183E	Q183E
		A701V	T1027I		R@213[b]	N440K	N501Y*	G339D	D614G	V213E	V213E	V213E	V213E
			V1176F		E@214	S477N*	Y505H*	S371F	H655Y	G252V	D253G	G339H	G339H
					V215P	T478K	D614G	S373P	N679K	G339H	G339H	R346T	R346T
					R216E	E484A*	H655Y	S375F	P681H	R346T	R346T	L368I	L368I
					G341D	Q493R*	N679K	T376A	Q954H	L368I	L368I	S371F	S371F
					R348K	Q498R*	P681H	D405N	N969K	S371F	S371F	S373P	S373P
					S371L	N501Y*	Q954H	R408S		S373P	S373P	S375F	S375F
					S373P	Y505H*	N969K	K417N		S375F	S375F	T376A	T376A
					S375F	D614G		N440K		T376A	T376A	D405N	D405N
					K417N	H655Y		L452R		D405N	D405N	R408S	R408S
					N440K	N679K		S477N*		R408S	R408S	K417N	K417N
					G446S	P681H		T478K		K417N	K417N	N440K	N440K
					S477N*	N764K		E484A*		N440K	N440K	V445P	K444R
					T478K	D796Y		F486V*		V445P	V445P	G446S	V445P
					E484A*	Q954H		Q498R*		G446S	G446S	N460K	G446S
					Q493R*	N969K		N501Y*		N460K	N460K	S477N*	N460K
					G496S*			Y505H*		S477N*	S477N*	T478K	S477N*
					Q498R*			D614G		T478K	T478K	E484A*	T478K
					N501Y*			H655Y		E484A*	E484A*	F486S*	E484A*
					Y505H*			N679K		F486S*	F486S*	F490S	F486S
					T547K			P681H		F490S	F490S	Q498R*	F490S
					D614G			N764K		Q498R*	Q498R*	N501Y*	Q498R*
					H655Y			D796Y		N501Y*	N501Y*	Y505H*	N501Y*
					N679K			Q954H		Y505H*	Y505H*	D614G	Y505H*
					P681H			N969K		D614G	D614G	H655Y	D614G
					N764K					H655Y	H655Y	N679K	H655Y
					D796Y					N679K	N679K	P681H	N679K

					N856K					P681H	P681H	N764K	P681H
					Q954H					N764K	N764K	D796Y	N764K
					N969K					D796Y	D796Y	Q954H	D796Y
					L981F					Q954H	Q954H	N969K	Q954H
										N969K	N969K		N969K
Envelope		P71L			T9I	T9I	T9I	T9I	T9I	T9I	T9I	T9I	T9I
										T11A	T11A	T11A	T11A
Membrane				I82T	D3N	A63T		Q19E		Q19E	Q19E	Q19E	Q19E
					Q19E			A63T		A63T	A63T	A63T	A63T
					A63T								
Nucleocapsid	D3L	L139F	P80R	D63G	P13L	R203K	P13L	P13L	P13L	P13L	P13L	P13L	P13L
	R203K	T205I	R203K	R203M	Δ31	G204R	Δ31	P151S	Δ31	Δ31	Δ31	Δ31	Δ31
	G204P	D402Y	G204P	G215C	Δ32	S413R	Δ32	Δ31	Δ32	Δ32	Δ32	Δ32	Δ32
	S235F	S413I		D377Y	Δ33		Δ33	Δ32	Δ33	Δ33	Δ33	Δ33	Δ33
					R203K		R203K	Δ33	E136D	R203K	R203K	R203K	R203K
					G204R		G204R	R203K	R203K	G204R	G204R	G204R	G204R
							S413R	G204R	G204R	A211S	S413R	S413R	S413R
								S413R		S413R			
NSP1			T78A			S135R	S135R	S135R	R124C	K47R	K47R	K47R	K47R
									S135R	S135R	S135R	G82D	G82D
												S135R	S135R
NSP2		T85I							Q376K				
NSP3	T183I	K837N	S370L	Q1216H	K38R	T24I	G489S	T24I	T24I	T24I	T24I	T24I	T24I
	A890D		K977Q	K1241R	Δ1265			G489S	V228A	G489S	G489S	G489S	G489S
	I1412T		A1183T	P1469S	L1266I				A231V			P1044L	
					A1892T								
NSP4			I23T	T492I	T492I	L264F	R222K	L264F		L264F	L264F	T83I	L264F
			T73I			T327I	T327I	T327I		T327I	T327I	L264F	T327I
						L438F		T492I		L438F	L438F	T327I	L438F
						T492I				T492I	T492I	L438F	T492I
												T492I	
NSP5		K90R		T196M	P132H	P132H	P132H	P132H	P132H	P132H	P132H	P132H	T93A
													K102R
													P132H
NSP6	Δ105	Δ105	Δ105	T77A	Δ105	F108L		Δ105	Δ105	Δ105	Δ105	Δ105	Δ105
	Δ106	Δ106	Δ106		Δ106			Δ106		Δ106	Δ106	Δ106	Δ106
	Δ107	Δ107	Δ107		Δ107			Δ107		Δ107	Δ107	Δ107	Δ107
	F108L	F108L	F108L		I189V			F108L		F108L	F108L	F108L	F108L
NSP7	[c]												
NSP8													

NSP9													
NSP10													
NSP11													
NSP12	R10G V11F P323L	P323L	P323L	P323L G671S	P323L	P323L	P323L	P323L	P323L	P323L G671S	P323L G671S	P323L G671S	P323L G671S
NSP13			E341D	P77L I334V		R392C	G287V R392C	R392C	M233I N268S R392C	S36P R392C	S36P R392C	S36P R392C	S36P R392C
NSP14				A504T	I42V	I42V		I42V	I42V	I42V	I42V	I42V I474V	I42V
NSP15						T112I	T112I	T112I		T112I	T112I	T112I	T112I
NSP16													
ORF3a		Q57H S171L	S253P	S26L		T223I	T223I	P@103 L106F T223I D@275	T223I	T32I T223I	T223I	T223I	T223I
ORF6						D61L		D61L		D61L	D61L	D61L	D61L
ORF7a				V82A T120I									
ORF7b								L11F					
ORF8			E92K	F120L									
ORF10		A8V											

The amino acid numbering is based on the "Wuhan" sequence

[a]Entries with a gray background in more than one variant

[b]Δ=deletion; @=insertion

[c]Blank rows denote no mutations found in that variant

*For spike protein these amino acids interact with ACE2 receptor of the host

As a function of time the spike protein has shown a significant number of synonymous and non-synonymous mutations over the course of the COVID-19 outbreak (Figure 5A). Note that in Figure 5A, each unique value of K_a and K_s (light orange and cyan respectively) are shown as dots, which correspond to a unique sequence of the protein, while the solid orange and cyan curves are based on the weighted average values for each day. The K_a/K_s ratio (red line in Figure 5B), the most commonly used metric, rose to a very high value in the initial months of the outbreak (January 2020) and has shown a tendency to decrease since then, with notable increases around July-September 2020 and May-July 2021. Note that during the initial period of the outbreak the data is substantially noisy due to the lower number of sequences, which affects the mean weighting (see Methods section for more details). The K_a/K_s ratio did not show any direct relationship with the number of new COVID-19 infections (dotted black lines). However, as reported previously by our group,[16] K_a showed a significant increase correlating with the Gamma/Delta variant related surge in 2021 and the Omicron surge in late 2021. A careful analysis of the data indicated that the non-synonymous mutations increased 10-14 days before each of the two surges. It should be noted that while the raw sequence data showed more unique sequences corresponding to synonymous mutations as compared to non-synonymous mutations until almost the end of 2021, the weighted non-synonymous K_a has showed a consistently higher value than the synonymous K_s since the onset of the Omicron variant in late 2021 (Figure 5A; note the difference in range between the cyan and light orange dots). The reason for this observation is that a significant number of sequences either closely resemble the Wuhan reference sequence or are weighted by more daily reports of low K_s values towards the beginning of the pandemic, while K_a is weighted more by the heavily mutated sequences that became more prevalent as time passed. Another possible explanation for this dramatic shift is that it could be a

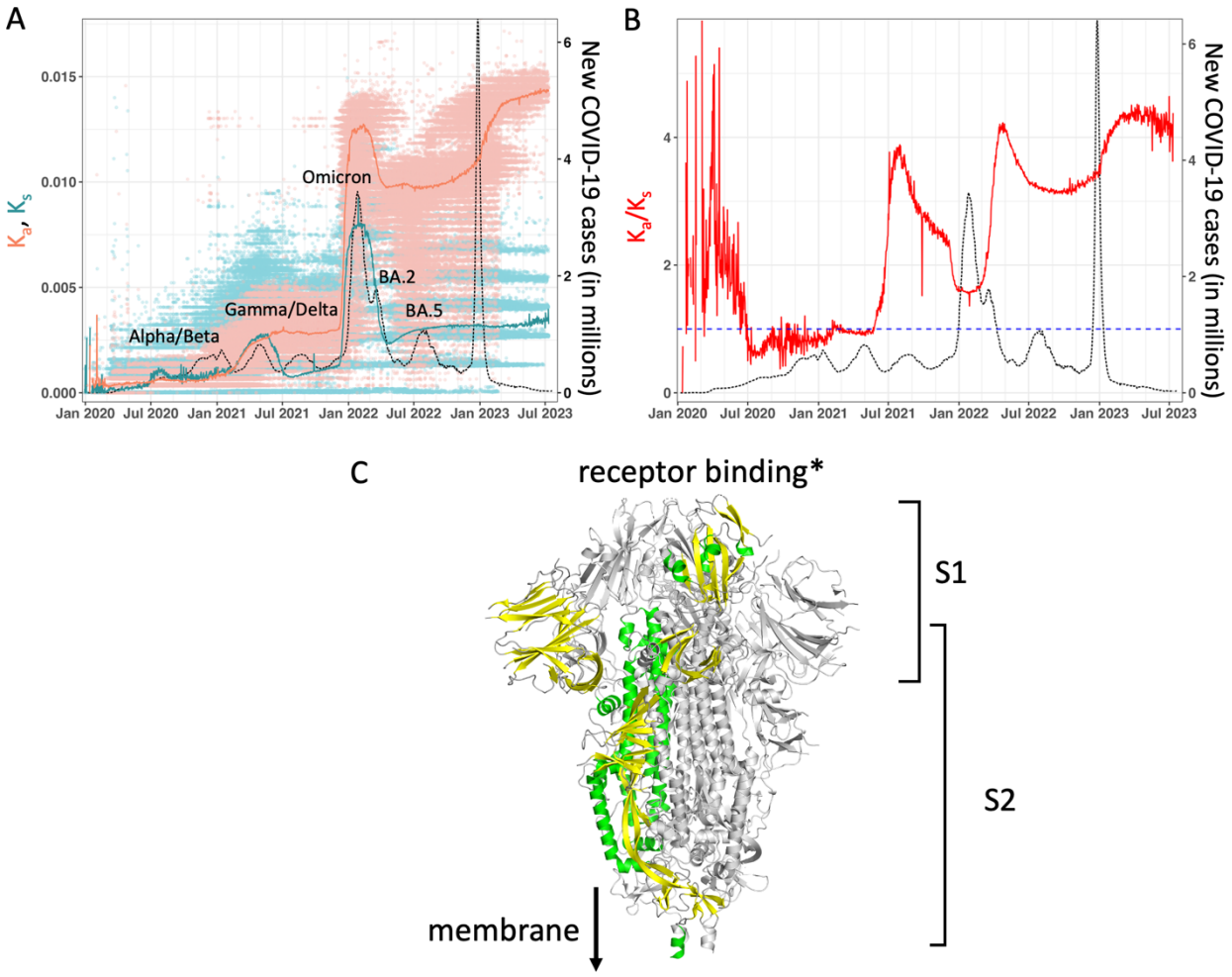

Figure 5: Spike protein mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) mutations are shown in light orange, while synonymous (K_s) mutations are shown in cyan. Individual dots correspond to unique sequences reported on different dates. The light orange and cyan curves are mean K_a and K_s values, weighted by number of sequences observed for each unique sequence. Note that a value of 0 corresponds to a sequence identical to the Wuhan sequence, and mean weighting for each unique sequence depends upon the number of identical reported sequences that match corresponding to a particular date. The dashed black curve shows new COVID-19 cases (sliding weekly average) across the world. The peaks for COVID-19 cases are labeled with prevalent variants. Alpha/Beta, Omicron, and Omicron BA.2 and BA.5 were the prevalent variants at the time of the labeled peaks. For the two peaks in 2021, the case was less clear, with Gamma and Delta variants being observed at different times in different parts of the world. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s) shown in red. As in panel A, the dotted black curve shows new COVID-19 cases. (C) Structure of the spike protein,

based on protein data bank (PDB) ID 6VXX.[17] The protein is a functional trimer, with one of them shown in green and yellow color, while the other two are colored in gray. The receptor binding region (corresponds to the entries marked by * in Table 1), and subunits S1 and S2 are marked. The direction of the domain that interacts with the membrane (not shown) is marked. Note, the missing regions in the structure are indicated by dashed lines.

result of the adaptive pressure from the mRNA-based vaccines targeting the spike protein which were widely deployed around this time. As noticeable in Figure 5A-B, both K_a and K_a/K_s are at their highest values as of July 2023, raising concerns about the long-term effectiveness of vaccines (and naturally acquired immunity) targeting the spike protein. The substantial number of mutations observed in Omicron XBB VOC further highlights this point. The data warrants a careful watch on the evolving situation. Note that there is large surge in COVID-19 cases between Dec. 2022 and Jan. 2023 (the largest number of COVID-19 cases reported during the course of the outbreak). The vast majority of these cases were reported from China. Information indicates that these were caused by the XBB variants, however, genomic information is not available in GenBank for these cases at the corresponding scale. Nonetheless, rise in K_a also corresponds to this surge in Figure 5A-B.

Envelope protein. This protein is essential for the production and release of mature viral particles.[18] This viroporin protein is produced in abundance throughout the course of infection, where it is thought to be most important during assembly, with a much smaller number of copies incorporated into mature virus particles.[19] It is also believed to bind toll-like receptors of the host cell, causing it to play a key role in the hyperinflammation present in COVID-19 cases.[20] More recent research suggests that ectopic expression of the envelope results in translational inhibition, as the viral protein binds to the initiation factor eIF2α.[21]

Amino acid substitutions for this protein are observed in the Beta variant and all the Omicron variants, which share the same mutation (Table 1). The mutation P71L, which is unique to the Beta variant, was found to have a slightly stabilizing effect on the protein structure, although its functionality is unknown[22] (Figure 6). The T9I mutation, which is shared across all Omicron variants, dramatically reduces the selectivity of the ion channel.[23] Due to this, when compared to the Wuhan type it will be less likely to kill the host cells and produce cytokine, which results in much less severe symptoms than variants lacking this mutation. The K_a/K_s plot is extremely noisy prior to the Omicron variant surges in late December 2021 (Figure 6B), and it shows a steep increase during this time and remains elevated to present day. This is also reflected in the K_a plot, which showed the most significant increase around the Omicron surge and further started increasing recently.

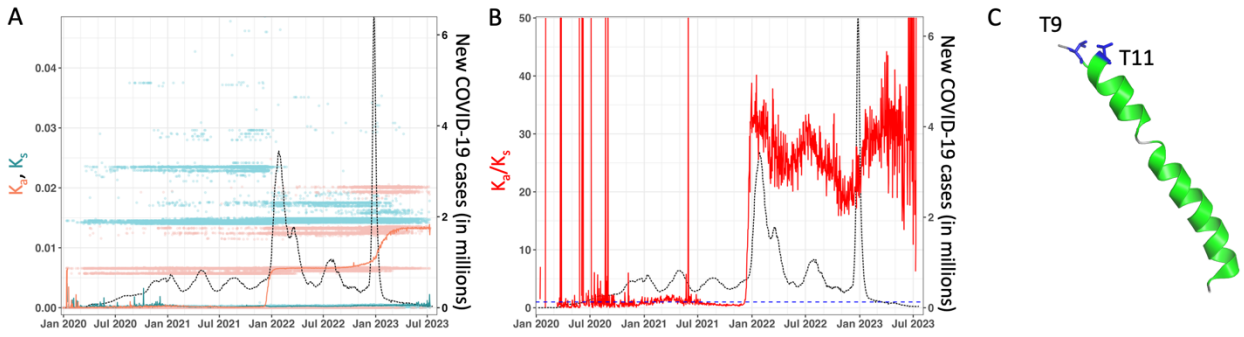

Figure 6: Envelope protein mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of the envelope protein, based on protein data bank (PDB) ID 7K3G.[24] The two residues T9 and T11 are shown, which show mutations since the Omicron VOCs. Note the residue 71 that showed mutation in Beta variant is located in the region where the situation information is not available.

Membrane protein. The membrane protein is believed to interact with the nucleocapsid protein in the formation of mature viral particles.[25] It is the most abundant protein in virions and has been found to be involved in a host of different functions, including but not limited to interfering with the host interferon system, inducing autophagy, serving as a protective antigen, and serving as a viroporin.[19]

Interestingly, for the membrane protein Q19E and A63T amino acid substitutions are observed in most of the Omicron subvariants, while the other variants, with the exception of the Delta variant, did not accumulate any mutations (Figure 7 and Table 1). The I82T mutant, which appeared in the Delta variant, has a slight stabilizing effect on the protein structure.[26] It is assumed that his mutation enhances glucose uptake during viral replication.[27] The K_a values show that amino acid substitutions started at the Delta variant surge in June 2021 and have remained high since then (Figure 7A). While the Beta variant's sequence shows a single amino acid substitution, it is not obvious on the K_a plot, likely due to how noisy the plot is at that region. The K_a and K_a/K_s plots show rise 2-4 weeks before the increase in number of infections for several VOCs (see Figure 7A-B). However, compared to other proteins, the K_a/K_s has shown a significant decrease recently.

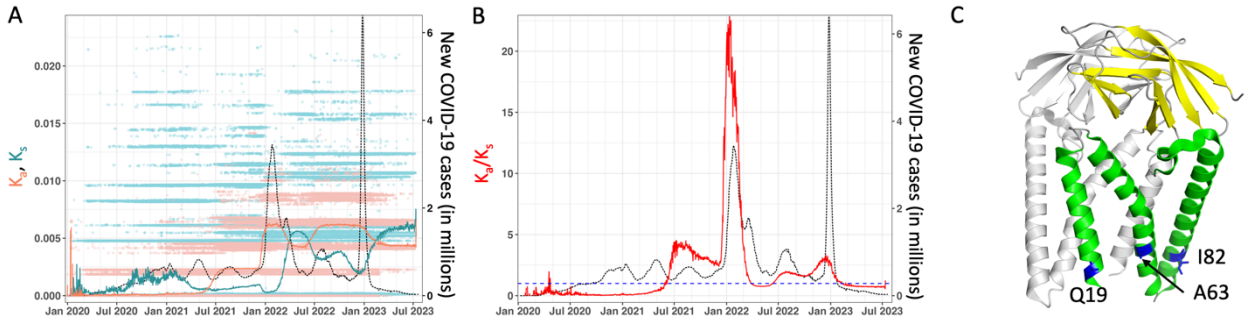

Figure 7: Membrane protein mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in

infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of the membrane protein, based on protein data bank (PDB) ID 8CTK.[28] The structure is shown as a dimer, with only one protomer colored in green and yellow. The location of three important residues showing mutations are marked.

Nucleocapsid. This protein is essential for packaging the viral genome into virion particles.[29] Specifically, this protein serves as the physical link between the positive-RNA genome and the envelope.[30] It has two domains; the N-terminal domain that binds to the viral RNA and the C-terminal domain that binds to the M-protein in the envelope with a large disordered region in the middle.[31] The other aspect of the nucleocapsid protein's function, an unknown, is its interaction with NSP3, suggesting a role in the viral life cycle.[31] A recent study has also correlated amino acid substitutions in this protein to the immune response and concluded that these mutations are correlated to the immune system evasion attempts, suggesting these sites as drug targets.[32]

Several mutations in this protein have been observed in all the variants (Figure 8 and Table 1). The double mutant R203K/G204R is present in all Omicron variants, with the Alpha and Gamma variants containing a similar double mutant, R203K/G204P. It was shown that this mutation augments the nucleocapsid phosphorylation, which in turn increases replication.[33] Interestingly, this mutation also increases the resistance to inhibition of the GSK-3 kinase. The Delta variant mutation G215C is located near the normally disordered region of the protein-protein interface (Figure 8C). Once introduced, this mutation causes the stabilization of the transient helix, enhancing the inter-protein interaction, which leads to the formation of a tetrameric rather than dimeric state. The nucleocapsid tetramer has a higher binding affinity to nucleic acid, which increases the virus' efficiency.[34]

From the broader genomic surveillance perspective, aside from the spike protein, the nucleocapsid protein demonstrates one of the highest propensities for mutation among the SARS-CoV-2 genes. Earlier in the pandemic, during the Alpha, Beta, Gamma, and Delta variant surges, the K_a/K_s ratio indicates that the protein was under positive selection, indicating some non-conservative amino acid substitutions (Figure 8B). However, starting at the Omicron subvariant surges, with the exception of the BA.2 subvariant, the K_a/K_s ratio switched to indicate a purifying pressure compared to the Wuhan strain. The K_a and K_s plots demonstrate the progression of the values though the surges. The ratio value remained larger than 1 until the Omicron BA.1 variant, when it dropped and remained low since.

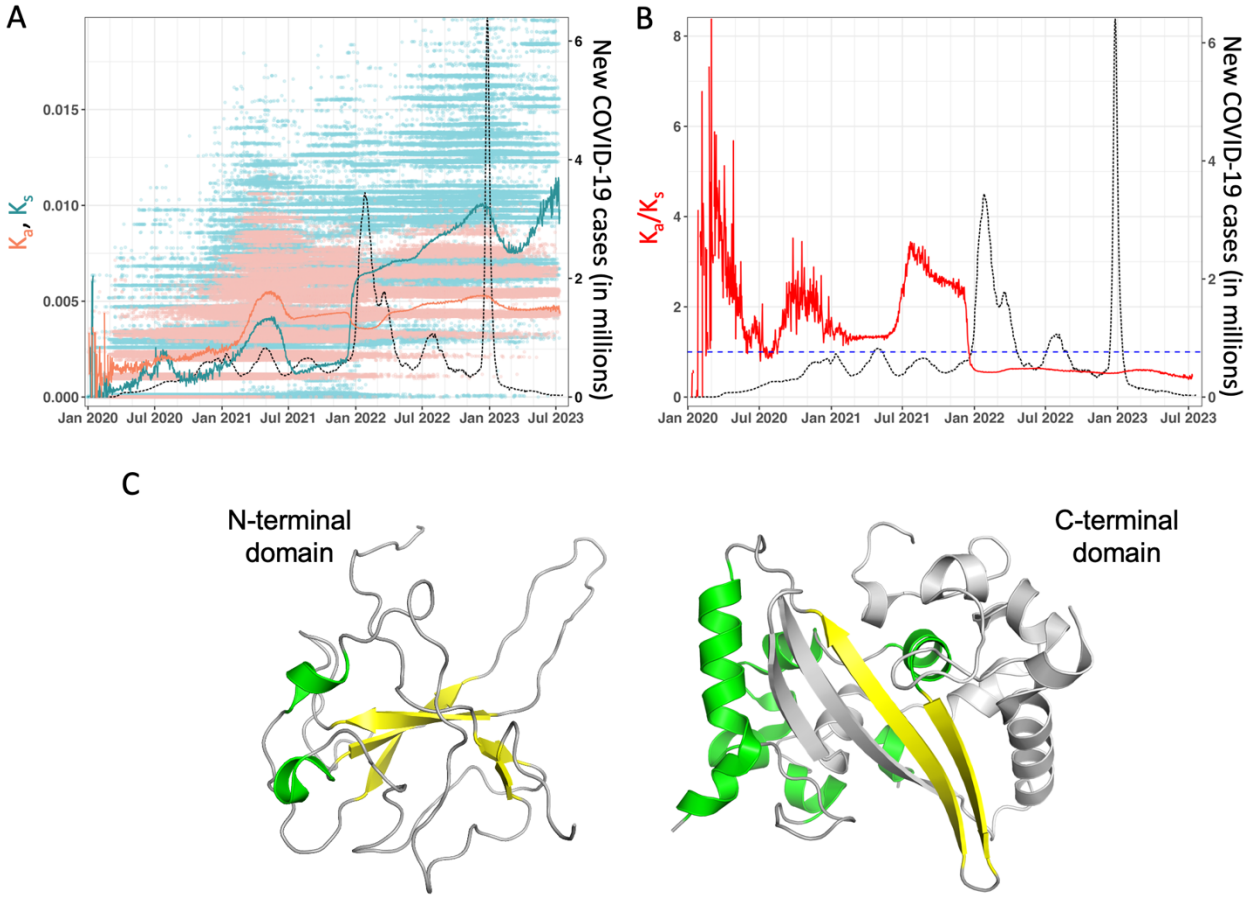

Figure 8: Nucleocapsid protein mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of the nucleocapsid protein, based on protein data bank (PDB) IDs 6YI3[35]for N-terminal domain and 7O05[36] for the C-terminal domain. The linker region between the two domains is not shown. The C-terminal domain structure is shown as a dimer, with only one protomer colored in green and yellow.

NSP1. Literature reports suggest that NSP1 is involved in repressing host transcription and preventing interferon induction.[37] Recent structural work has proposed its mechanism of action involves the last 33 amino acids of the C-terminal domain of the protein tightly binding to the 40S ribosomal subunit, physically blocking mRNA from entering the channel and therefore preventing translation.[38-40] It has further been hypothesized that NSP1 skews kinetics in its favor by uncapping cellular mRNA, leading to its degradation, while also recognizing a viral 5' UTR stem loop to override the translation block.[41] Moreover, NSP1, along with NSP14, has been shown to inhibit the accumulation of the human long interspersed element 1 (LINE-1) open reading frame ORF1p, suggesting a direct interfering with the interferon response.[42]

Figure 9 shows the summary of mutation behavior for NSP1 over the course of the COVID-19 outbreak. The K_a and K_s values show typical trends of increases over time. However, no correlation between mutation values and increase in number of new COVID-19 infections is particularly observable for NSP1. K_s did show a slight increase around April 2021; however, no significant increase was observed in K_a values in the initial period of the pandemic. This signals nucleotide-level changes in NSP1 sequences, with silent mutations that cause no changes in the amino acid sequence at the protein level. K_a and the K_a/K_s ratio did show a significant increase after the Omicron BA.1 surge (January 2022 onwards). This increase coincides with the most

common S135R amino acid substitution observed in all Omicron variants except BA.1 (Table 1). Not much is known about the functional importance of this and other mutations. Structural information is only available from residues 13 to 127 (Figure 9C), with observed mutations shown. Note that S135R falls outside the known structure.

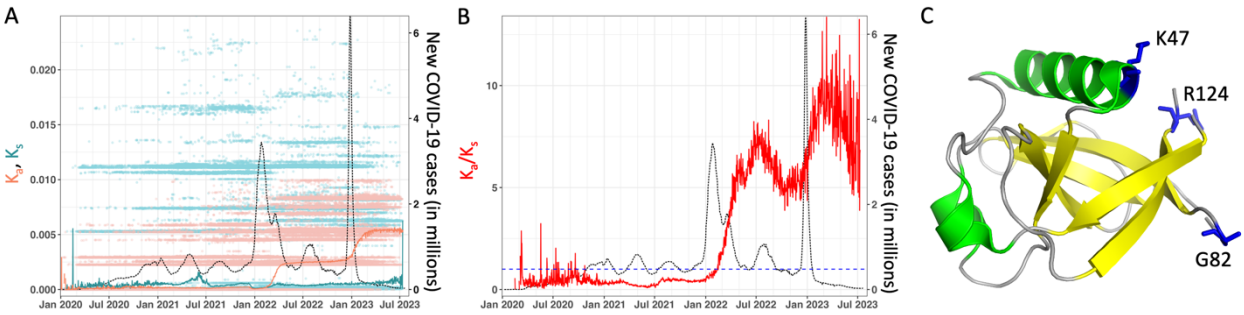

Figure 9: NSP1 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP1, based on protein data bank (PDB) ID 7K3N.[43] Location of residues showing mutations are marked, note S135 is not listed as it falls outside the region 13-127 for which structural information is available.

NSP2. Research suggests this protein might be involved in viral genome replication.[44] This is supported by a study finding NSP2 to be found in the perinuclear foci and co-localized with the nucleocapsid protein, suggesting a presence at the site of viral RNA replication.[45] More recently, a function involving the inhibition of miRNA silencing pathways has been suggested, allowing for NSP2 to suppress the host immune system through taking control of posttranscriptional silencing.[46]

Examination of the amino acid alignments of the major variants showed single, unique mutations in the Beta and Omicron BA.5 variants (Figure 10 and Table 1). One such mutation, T85I, is located on the protein's surface (Figure 10C). The substitution of a hydrophobic residue

in the place of a polar residue indicates a potential for destabilization of the structure. Indeed, in accordance with a computational study, the T85I mutant is more flexible and less stable than the Wuhan-type protein.[47] Another mutation observed in Omicron BA.5 was Q376K, however, this mutation was not observed in other VOCs. Despite the noisy nature of the K_a/K_s plot, the K_a/K_s ratio decreases sharply after January 2021, around the end of the beta surge (Figure 10B).

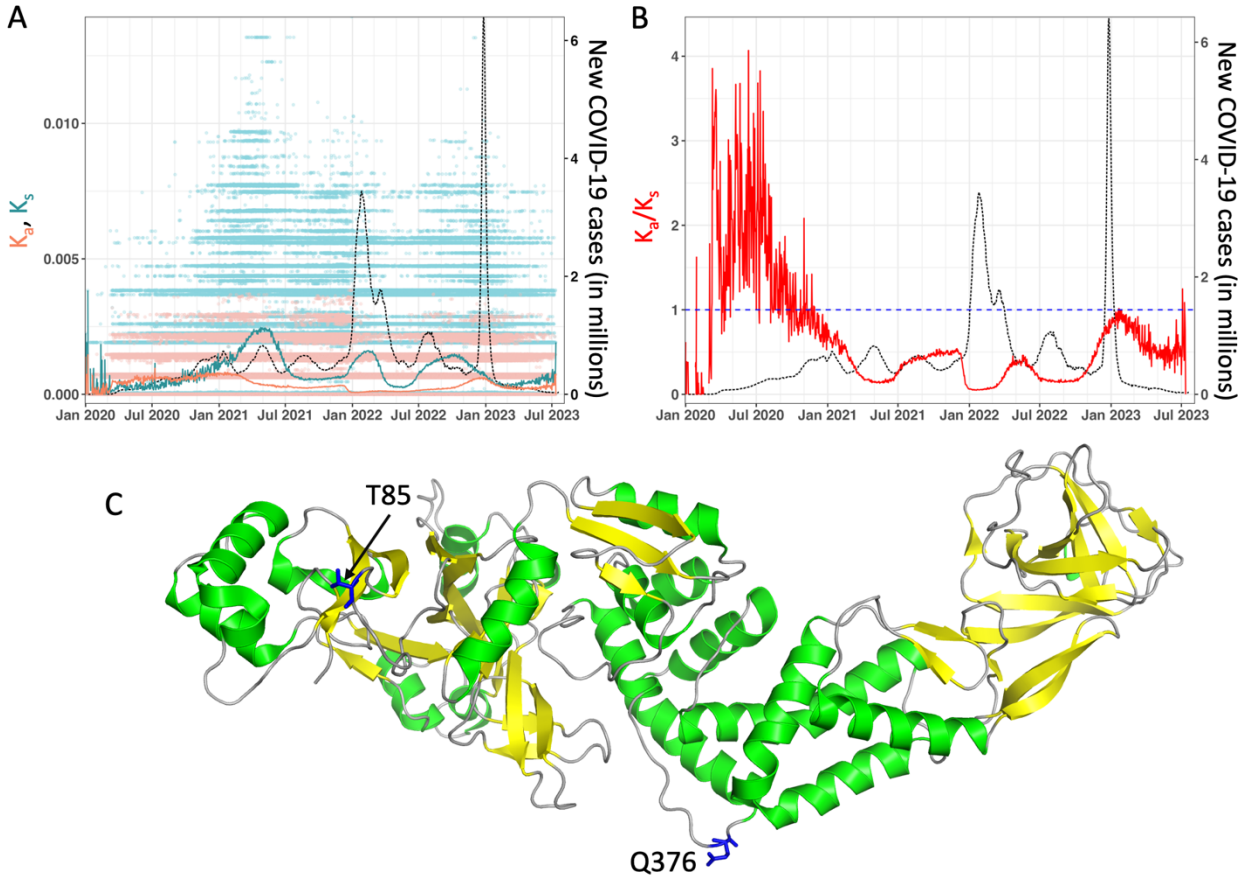

Figure 10: NSP2 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP2, based on protein data bank (PDB) ID 7MSW [DOI:10.2210/pdb7MSW/pdb].

NSP3. This is the largest protein in SARS-CoV-2 with multiple domains (see Figure 11). Research has shown that it interacts with the nucleocapsid protein.[31] However, the function of this interaction is still not fully understood. It is also speculated that this protein, along with NSP5 and NSP14, might be involved with maintaining an optimal environment for viral RNA synthesis, as research demonstrated their ability to repress the expression of the reverse transcriptase activity of the LINE-1 open reading frame ORF2p.[42] Additionally, NSP3 pairs with NSP4 in the formation of replication organelles for the virus.[48] Kahn et al have shown that residues 499 to 533 on NSP3 interact with the nucleocapsid protein, making them a potential target for drugs.[31] Supporting this potential usage, none of those residues seem to be substituted in any of the variants, suggesting that they are essential for its function (Table 1). Additionally, NSP3 is a papain-like cysteine protease (PLPro) that processes the viral polyprotein.[49] There is active research to develop inhibitors for this protein; specifically, targeting its conserved macro domain (Mac1), as it is critical for pathogenesis of SARS-CoV-2.[50]

Amino acid substitutions lack a general pattern observed throughout all the VOCs, with many variants containing unique mutations. The substitutions T24I and G489S are the only mutations to occur more than once, with concurrent occurrence in all variants except Omicron BA.5 beginning with Omicron BA.4. The K_a data is reflective of the adaptational mutations we observed (Figure 11A). At the start of March 2021, the average K_a values of the isolates were higher for the Gamma and Delta variants. The values started to decrease in December of 2021, but still remained higher than the background.

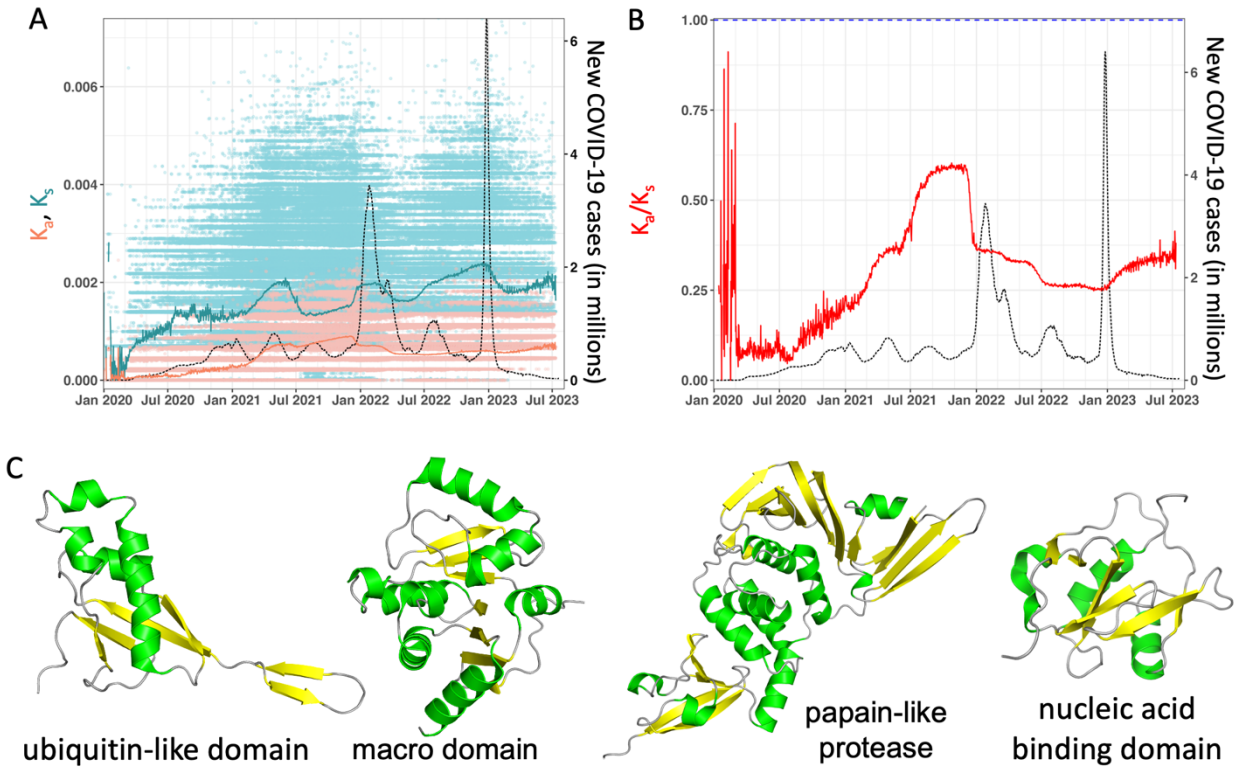

Figure 11: NSP3 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP3, based on protein data bank (PDB) 7KAG [DOI:10.2210/pdb7KAG/pdb] (ubiquitin-like domain, Ubl1), 6YWK (macro domain, Mac1),[52] IDs 7CMD (papain-like protease, PLpro),[51] and 7LGO [DOI:10.2210/pdb7LGO/pdb] (nucleic acid binding domain).

NSP4. This gene encodes an endoplasmic reticulum-bound protein.[54] Comparative studies suggest that this protein may play a role in the virus replication assembly process.[55] Specifically, NSP4 has been linked to intracellular double-membrane vesicle formation, with NSP4 working in conjunction with NSP3 to perform membrane pairing.[56] More recently, it has been shown that the endoplasmic reticulum proteins VMP1 and TMEM41B are essential, as they facilitate NSP3 and NSP4, in conjunction with the ER, to generate the replication organelles.[48]

No substitutions were observed in the Alpha, Beta, and Omicron BA.5 variants, but amino acid substitutions were observed in all other VOCs (Table 1). Interestingly, the Delta and Omicron BA.1, BA.2, and BA.4 variants all share the T492I substitution, making it the most common substitution. This residue is located in the M-domain, within the lipid bilayer. Note, structural information is not available for NSP4 from SARS-CoV-2. Due to the increase in hydrophobicity caused by this mutation, it is suggested that this mutation stabilizes the RTC, enhancing replication.[57] The T327I mutation, found in Omicron variants BA.2-BA.4, is suspected to function similarly. The K_a/K_s plot suggests an increase in non-synonymous mutations around January 2022 (Figure 12B). This plot, in conjunction with the separate K_a and K_s plots, suggest that the number of nonsynonymous substitutions are increasing, starting around the Omicron BA.1 surge. The observed amino acid substitutions echo the plots' trajectories. It seems that the Omicron subvariants mark the beginning of retaining the same core set of substitutions.

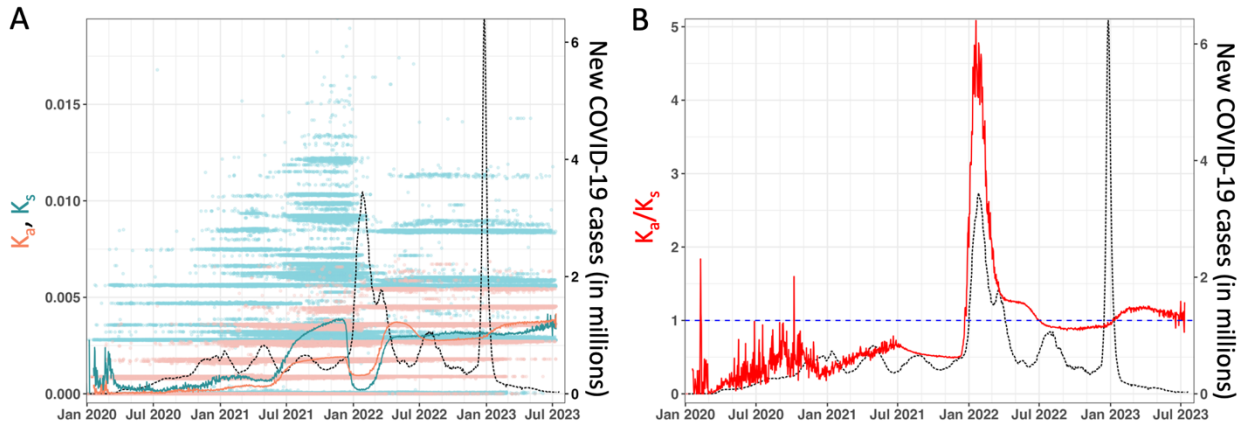

Figure 12: NSP4 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). As in panel A, the red curve shows new COVID-19 cases.

NSP5. This protein is a 3C-like proteinase. It processes the polyproteins PP1a and PP1ab into their individual constituent proteins.[58] It also is believed to play a role in evading the host's immune system, as it also cleaves NF-kB essential modulator (NEMO) at multiple sites, ultimately resulting in inhibition of the production of interferon-beta (IFN-b).[59] Recent work suggests that NSP5 also works with NSP3 and NSP14 to repress the expression of the reverse transcriptase activity of the LINE-1 open reading frame ORF2p.[42]

Five unique amino acid substitutions were observed in this protein (Figure 13 and Table 1): K90R, which is observed only in the Beta variant; T196M, which is observed only in the Delta variant; T93A and K102R, which are observed only in the Omicron XBB.4 variant; and P132H, which is observed in all Omicron variants. No amino acid substitutions were observed in the Alpha or Gamma variants. The K90R mutation, located in domain I of the protein, has shown functional relevance despite both residues containing polar, positive sidechains (Figure 13C). In one study, the catalytic efficiency of this mutant is 56% that of the Wuhan type.[60] The P132H mutation, located in domain II, displays a different effect. Despite a 2.6ºC lower thermal stability,[61] the enzymatic activity is 44% higher than that of the Wuhan strain.[60]

The plots, and particularly the data for K_a, demonstrate two distinct regions of increase in the nonsynonymous amino acids' substitutions (Figure 13). The first region spans from the start of the pandemic until February 2021; this region includes the both the Alpha and Beta variant surges. The second region begins around mid-December 2021 and runs until present day. Analyzing the amino acid substitutions confirm that all the Omicron variants have both higher nonsynonymous and synonymous substitutions. The nonsynonymous mutations can be seen clearly at the nucleotide sequence level.

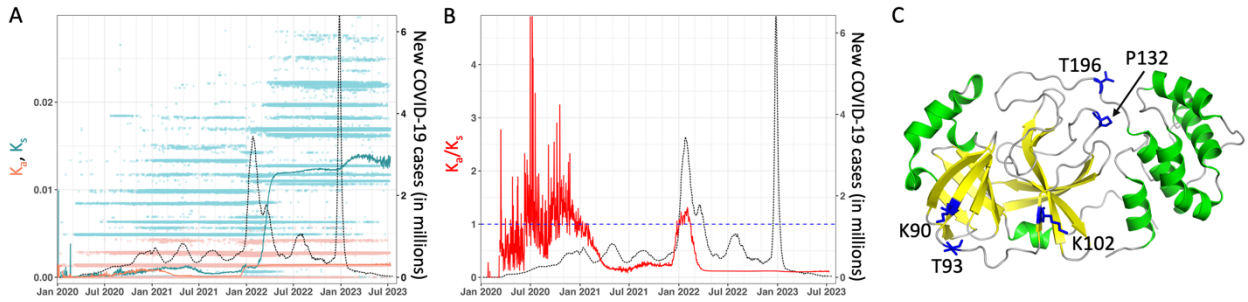

Figure 13: NSP5 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP5, based on protein data bank (PDB) ID 7LYH.[62]

NSP6. This protein is an endoplasmic reticulum membrane protein that regulates autophagy of the viral particles, which ultimately allows the viral particles to be successfully exported outside the host cell.[63] Hence, NSP6 is an essential part of protecting viral particles. In corroboration with this, a study correlated a single nucleotide polymorphism, the substitution L37F, to destabilization of an NSP6 fold that results in decreased virulence.[64] Indeed, recent research suggests that NSP6 and the spike proteins are essential for the Omicron subvariants' attenuation, which further suggests its role in evasion of the immune system.[65] Remarkably, it has been shown that NSP6 carries different sets of mutations based on geographical locations.[66]

Interestingly, the studied L37F substitution was not observed in any of the variants of concern; however, other changes were observed (Table 1). Three-amino acid-long deletions at positions 105, 106, and 107 were observed in all variants except for the Delta and Omicron BA.2, BA.3, and BA.5 variants. Additionally, three additional amino acid substitutions were observed in different variants, with the F108L mutation shared by several different variants. The structure of NSP6 from SARS-CoV-2 is not available. The K_a and K_a/K_s plots suggest a high number of nonsynonymous substitutions February 2021, around the gamma surge, and

December 2021, at the Omicron surge (Figure 14B). It also displays a lower K_a/K_s ratio around August of 2021, coinciding with the Delta variant surge. This trend is also reflected in the dearth of amino acid substitutions in Delta variant.

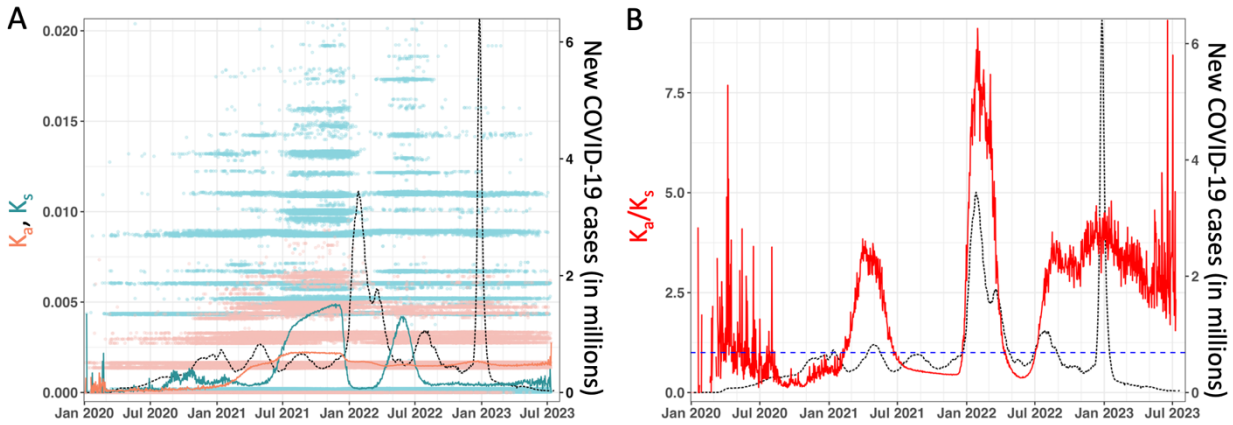

Figure 14: NSP6 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s).

NSP7. This protein is a cofactor of the replication complex, which includes NSP12 (RNA-dependent RNA polymerase) and NSP8 proteins.[67] Hence, it plays an essential role in viral genome replication and transcription. It most closely interacts with NSP8, which it forms a super-complex with, that is a critical cofactor for NSP12.[68] Interestingly, no amino acid substitutions have been observed in any of the variants (Table 1). This suggests that this gene is under little to no adaptation pressure, possibly due to its high interactivity with other viral proteins. The K_a and the K_s plots show that there are no nonsynonymous amino acid substitutions of note (Figure 15).

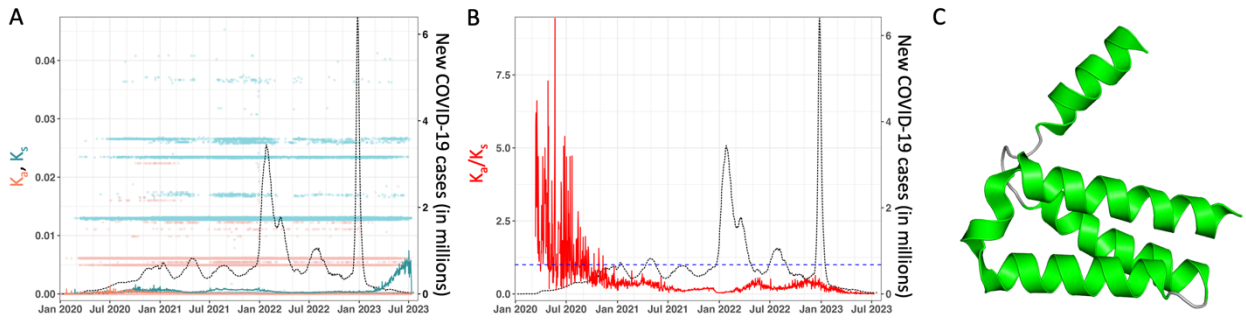

Figure 15: NSP7 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP7, based on protein data bank (PDB) ID 7JLT.[69]

NSP8. This protein, along with NSP7, is a cofactor for the NSP12 enzyme. Studies suggest that the transcription complex is a hetero-tetramer composed of one molecule of NSP7, one molecule of NSP12, and two molecules of NSP8.[67] Like NSP7, no amino acid substitutions were observed in any variant, suggesting that those proteins are under little to no pressure to adapt (Table 1). Additionally, the same observation of the mutation pattern is noted in the individual and K_a/K_s plots for this protein (Figure 16). Of note, the K_s plot suggests a sharp increase in synonymous nucleotide substitutions, as it shows an increase after the Omicron BA.2 surge after April 2022.

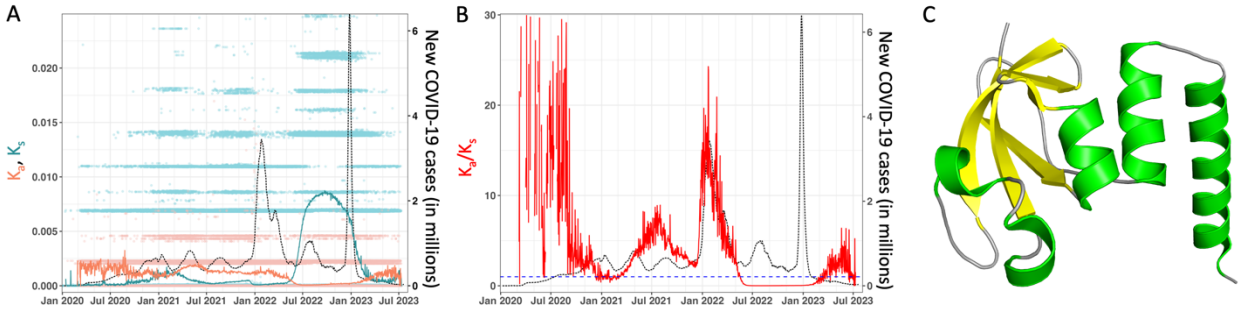

Figure 16: NSP8 mutation summary over the course of the COVID-19 outbreak. ((A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other

details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP8, based on protein data bank (PDB) ID 7JLT.[69]

NSP9. This protein is believed to reside in the endoplasmic reticulum.[70] Several research efforts suggest that it interrupts host cell protein trafficking.[71] It also is reported that NSP9 sequesters the NUP62 protein, a part of the nuclear pore complex, affecting the transportation of different components of the host's immune response.[70] More recently, it has been shown that this protein interacts with the E3 RING ligase enzyme, which is part of the ubiquitin system that processes different cellular functions, including immune signaling.[72] This interaction might be part of the successful evasion of the host immune system. Like NSP7 and NSP8, amino acid substitutions have not been observed in any of the variants (Table 1). The K_s plot shows that the average K_s increases after March 2021 at the Gamma variant surge and starts to decrease again after August 2021 (Figure 17). K_s values dramatically increase again and stabilize at that elevated level after January 2022. These changes in K_s values are echoed in the nucleotide sequences in the Gamma and Omicron BA.2 – BA.5 variants.

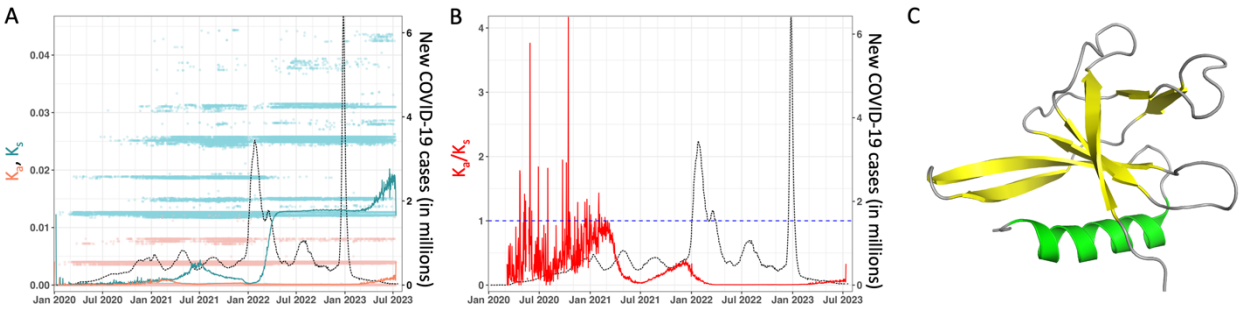

Figure 17: NSP9 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP9, based on protein data bank (PDB) ID 7BWQ.[73]

NSP10. This protein is an essential component in activating the replication/transcription complex. It is an activator for the error-repair enzyme 3'-5' exonuclease (NSP14) enzyme.[74,75] Another important function attributed to this protein is its interaction with NSP16 in order to activate its 2'-O-methyltransferase activity to evade the host immune system.[75,76] This gene has not displayed any amino acid substitutions throughout the outbreaks (Table 1). The K_s plot shows a significant increase around December 2021, which continues until a decrease in March 2022, coinciding with the Omicron BA.1 surge (Figure 18). This is confirmed when comparing the nucleotide alignments of the Omicron BA.1 variant against the Wuhan variant. Despite these nucleotide changes, the K_a plot suggests that there are no amino acid substitutions.

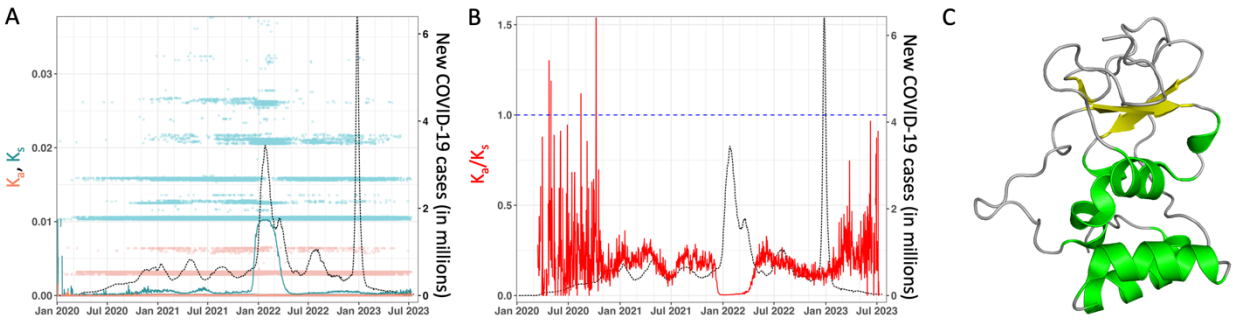

Figure 18: NSP10 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP10, based on protein data bank (PDB) ID 7MC5.[77]

NSP11. Research suggests that this is an intrinsically disordered protein (IDP), hence it lacks an ordered structure; however, in some solvent conditions mimicking membrane environments (e.g., TFE, SDS) it forms an α-helix, suggesting that it may have some membrane-related role.[78] As the smallest of the encoded genes, it contains just 13 amino acids. There have been no amino acid substitutions in any of the VOCs (Table 1). Due to the protein's small size and resulting

small database depth, the plots are extremely noisy; however, both individual and ratio plots show that K_a and K_s have been stable throughout all the surges (Figure 19).

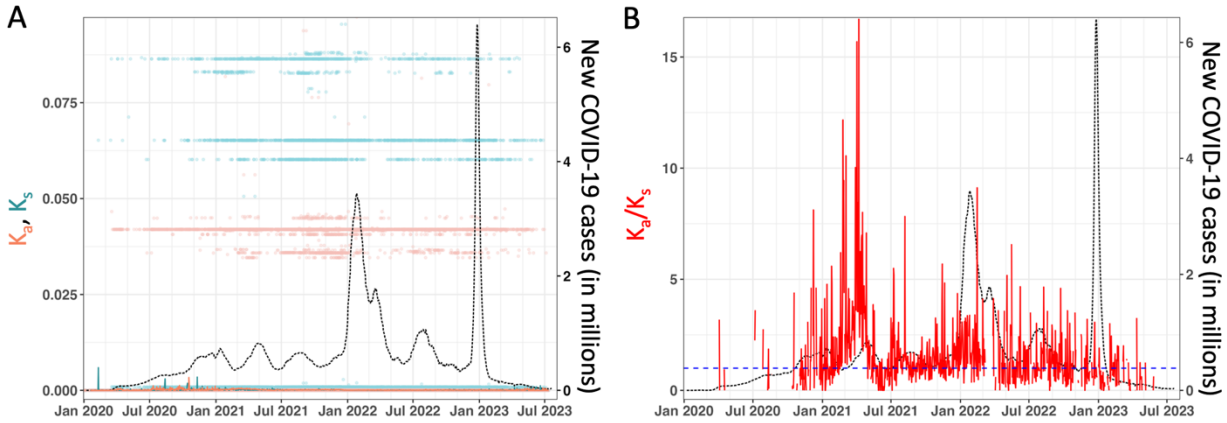

Figure 19: NSP11 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s).

NSP12. This is the catalytic subunit of the RNA-dependent RNA polymerase (RDRP). This catalytic subunit complexes with one molecule of NSP7 and two molecules of NSP8 and is responsible for the replication of the viral genome.[68] Recently, a study suggested the use of modified RNA templates as a drug to inhibit NSP12.[79]

All variants share a P323L mutation when compared to the Wuhan genome (Figure 17 and Table 1). The P323L substitution is located in the RDRP interface domain. This mutation increases the stability of the RDRP structure, while simultaneously weakening the RDRP-NSP8 interaction.[80] It is suggested that both stabilizing the RDRP structure and destabilizing its interaction with the co-factors reduces the proofreading capability, significantly increasing the RNA replication efficiency along with error-derived mutation rate.

The K_a/K_s plot echoes these observations (Figure 20B). Note that the plot decreases during the Gamma variant surge between mid-April to mid-May of 2021, then immediately increases again at the Delta variant surge between June and November of 2021. It dramatically decreases again at the Omicron variant surges. While the nonsynonymous mutations are somewhat limited in scope and diversity, there are synonymous substitutions throughout the variants.

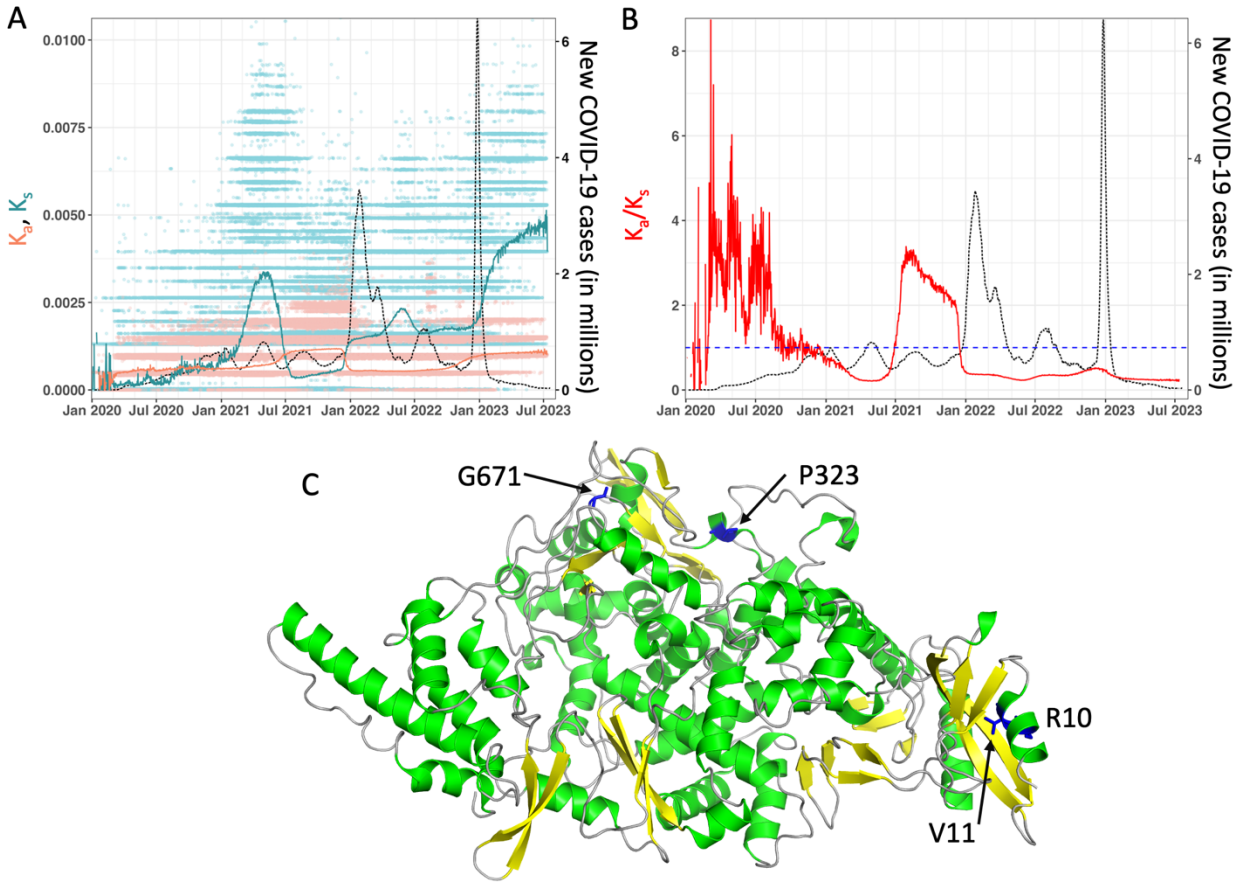

Figure 20: NSP12 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP12, based on protein data bank (PDB) ID 7KRN.[81]

NSP13. This protein is the helicase part of the mature replication complex.[82] There are a couple of roles attributed to this protein. One role is for proper folding of RNA genome, while the second role is as a potential suppressor of interferon I productions.[82]

Amino acid substitutions occur in all but the Alpha, Beta, and Omicron BA.1 variants, with the Omicron BA.2-BA.5 variants sharing the same R392C substitution (Figure 21 and Table 1). In accordance with molecular dynamics simulations, the Gamma variant-limited E341D mutation and Delta variant-limited P77L mutations demonstrate a higher binding affinity to the TBK1 than the Wuhan strain, thus more effectively evading the immune response.[83] The shared Omicron variants' mutation, R392C, is located close to the active site of the Rec1A domain. Although the effect on efficiency is yet to be determined, this mutation increases the flexibility of the protein, thus decreasing the stability.[84] The K_a and K_a/K_s plots reflect these nonsynonymous mutation observations (Figure 21 A-B). The average values of the K_a/K_s ratio increase around the Gamma and Delta variant surges between March and September of 2021. It subsequently decreases again, until the Omicron BA.2 variant surge emerges around March of 2022. It has stayed elevated since then.

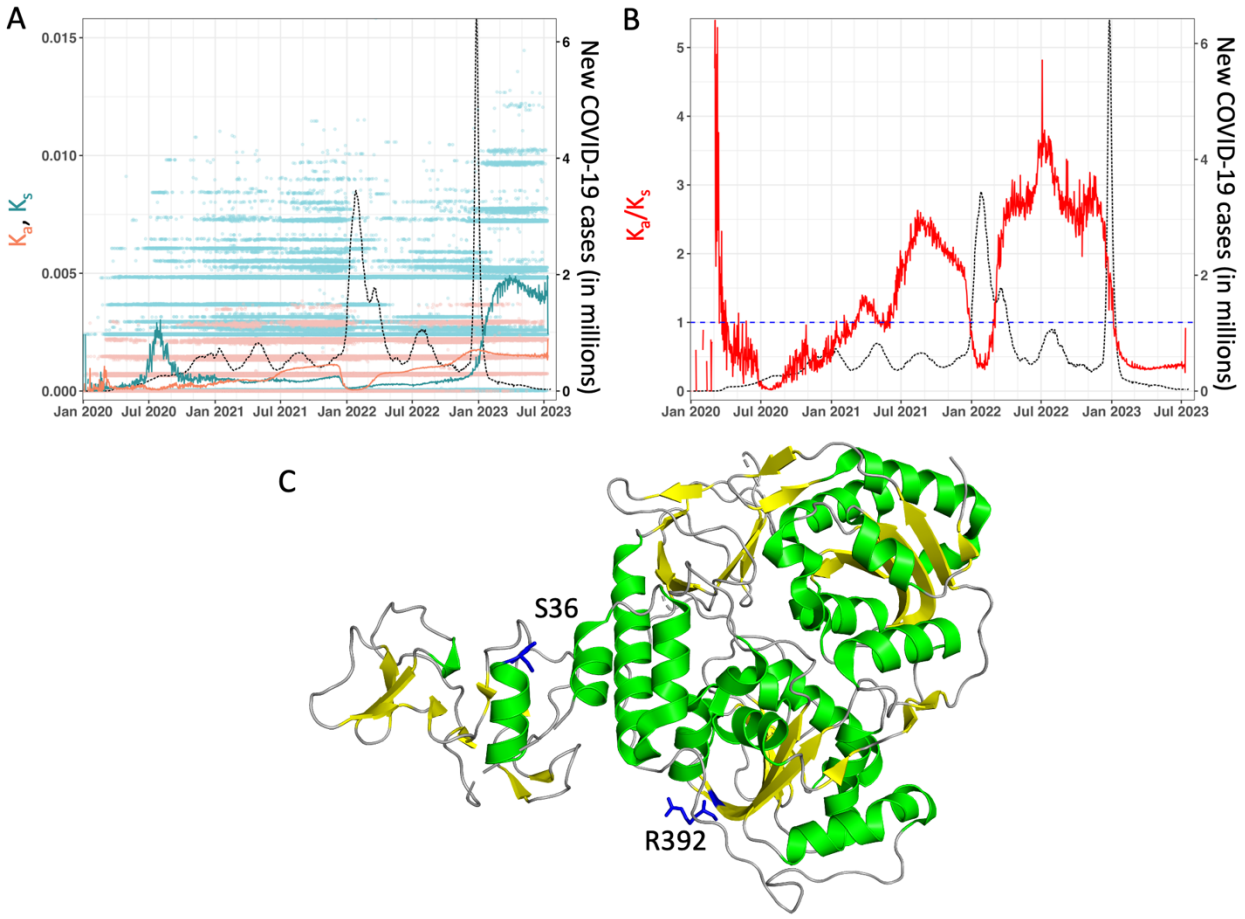

Figure 21: NSP13 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP13, based on protein data bank (PDB) ID 5RL6.[85] Note, the missing regions in the structure are indicated by dashed lines.

NSP14. This protein is also part of the replication complex, where it acts as the 3' to 5' exonuclease part of the complex.[86] It ensures an accurate viral transcription by excising mismatched bases. Additionally, this enzyme has a second activity, which is as a C-terminal N7-methyltransferase.[87] One of the clinically relevant roles attributed to NSP14 is interferon I response inhibition.[88] The protein contains two unique substitutions, one of which is shared by

the Omicron BA.1, BA.2, and BA.4 variants (Figure 22 and Table 1). Interestingly, all the observed substitutions are conserved hydrophobic substitutions. The K_a/K_s plot shows that there is an increase in nonsynonymous amino acids substitutions starting at the Delta variant surge around June of 2021. The plot shows a continuous increase during the span of the entire Omicron subvariants surges.

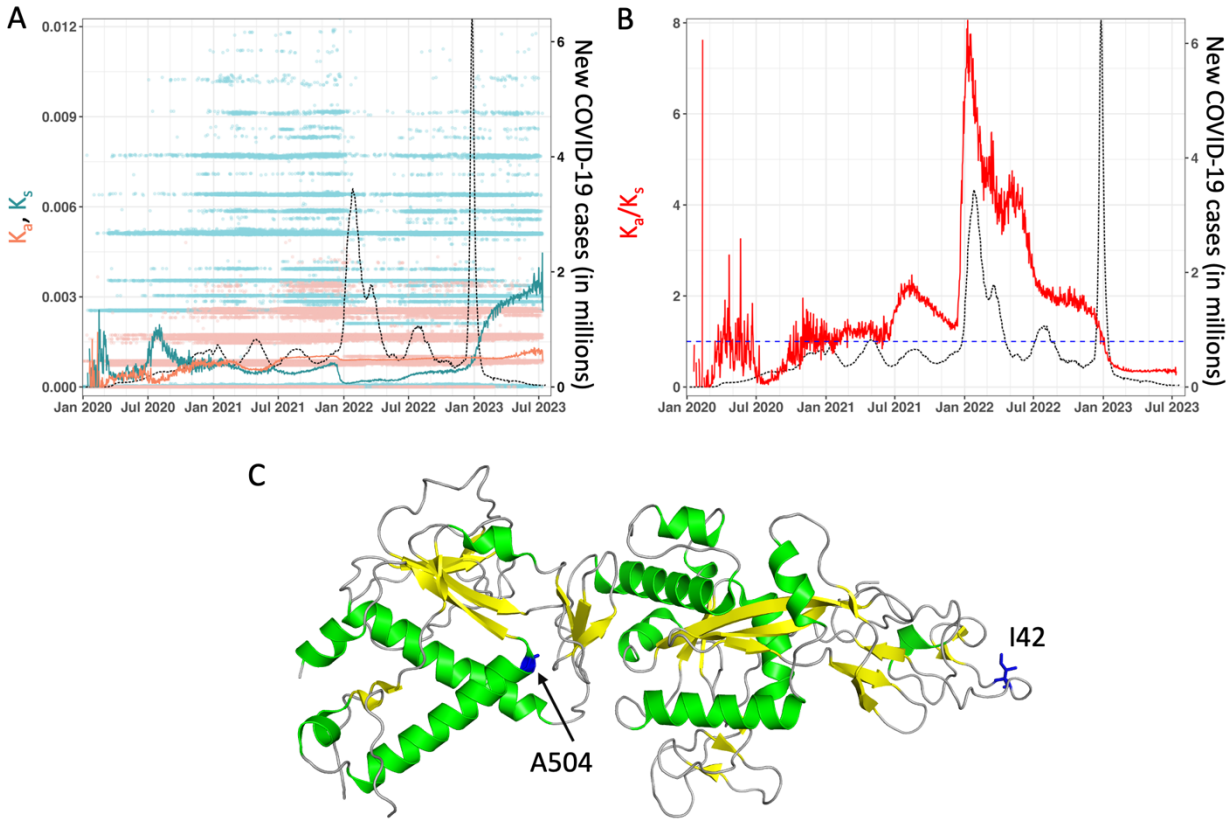

Figure 22: NSP14 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP14, based on protein data bank (PDB) ID 7N0D.[89]

NSP15. This protein is a conserved endoribonuclease that degrades pyrimidines at the 3' end with specificity to uridine.[90] It preferably degrades unpaired pyrimidines.[90] The activity of NSP15 is affected by several different factors, including RNA secondary structure. The presence of a 2'-O-ribose methyl group in the RNA inhibits the NSP15 activity.[91] Interestingly, NSP16, which is encoded immediately downstream of it, is predicted to be a 2'-O-ribose methyl transferase. This has led to the hypothesis that NSP16 is a regulator of NSP15.[92] The function of NSP15 is to degrade any dsRNA intermediates, thus preventing the recognition of the viral genome inside the host cell. This ultimately aids in evasion of the host immune system by delaying the type I interferon response.[93] More interestingly, NSP15 is unique to the *nidovirus* family, with no orthologues in humans, which makes it an attractive drug target.[94] Recently, research has shown that NSP15 binds to the E3 ligase RNF41, which suggests a disruption of the immune system.[72]

The amino acid substitution T112I is the only mutation observed (Figure 23 and Table 1). This substitution seems to be Omicron-specific, appearing in all subvariants except BA.1 and BA.5. This substitution does not match any known vital amino acid position in this protein. The plots show little mutational activity, with the K_a/K_s plot dropping its small elevation at the onset of the Omicron variant in late December 2021. After the beginning of the Omicron variant surges beginning in 2022 both the K_a and K_s values remain at a steady elevated level.

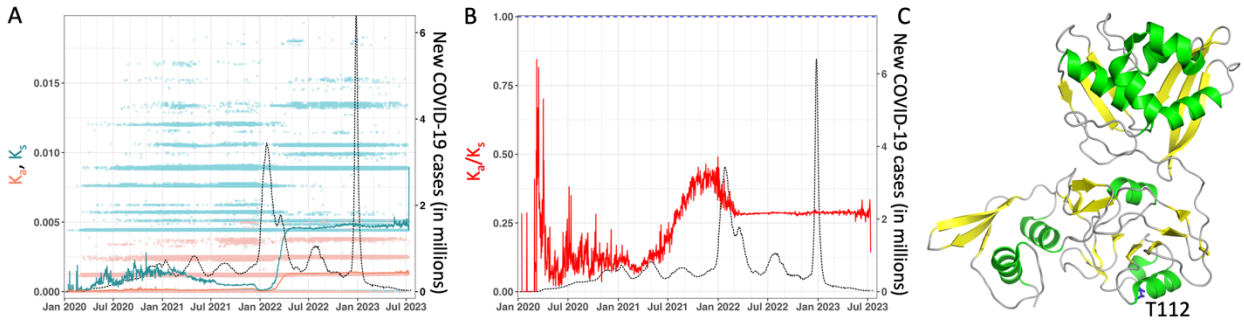

Figure 23: NSP15 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP15, based on protein data bank (PDB) ID 7ME0 [DOI:10.2210/pdb7ME0/pdb].

NSP16. This protein is a 2'-O-methyltransferase. It has been hypothesized that it regulates NSP15.[92] This protein also plays an essential role in immune system evasion by mimicking the activity of its human homolog, cap-specific mRNA (nucleoside-2'-O-)-methyltransferase (CMTr1), methylating mRNA to both improve translation efficiency as well as camouflaging mRNA to allow it to remain undetected from intracellular pathogen recognition receptors.[95] Additionally, it needs to interact with NSP10 in order to be activated.[76]

No substitutions were observed for this protein in any of the major variants (Table 1). In accordance with the sequence analysis, the K_a, K_s, and K_a/K_s plots show no noticeable change throughout the surges (Figure 24).

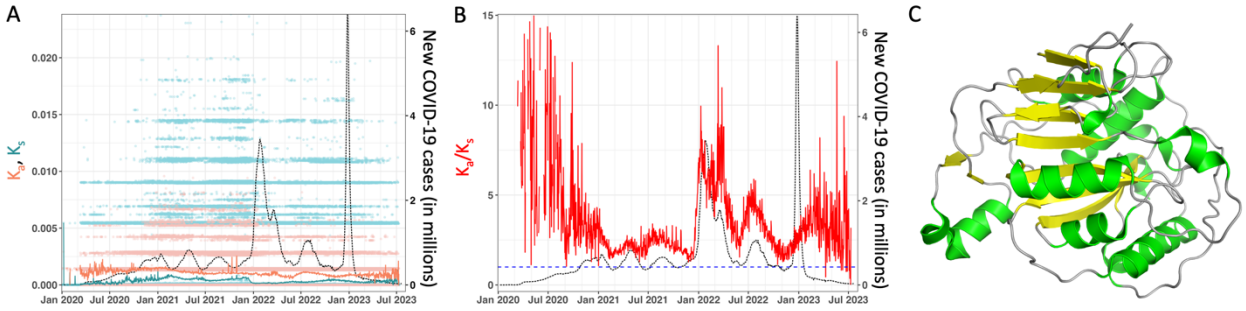

Figure 24: NSP16 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of NSP16, based on protein data bank (PDB) ID 6WKS.[96]

ORF3a. This protein is an integral membrane protein. It potentially is involved with the phases of the viral life cycle and viral genome replication and release.[97] It helps to amplify viral release by promoting lysosomal exocytosis.[98] Recently, research has shown that his protein binds to the E3 ligase TRIM59.[72] Additionally, ORF3a, in conjunction with ORF7a, has demonstrated a reduction of antigen presentation by the human major histocompatibility complex (MHC-II), thus increasing the viral evasion of the host's immune system.[99]

A number of different amino acid mutations occur in all variants except the Alpha and Omicron BA.1 variants (Figure 25 and Table 1). With the exception of the T223I substitution, these mutations are unique for each variant. The mutation Q57H, found in the Beta variant, was initially predicted to bind to impact binding to the spike protein, causing more severe disease;[100] however, experimental work was found that this mutant has nearly the same activity as the Wuhan ORF3a.[101] The Omicron BA.2 substitution L106F is located near the ion channel, although it is considered to be on the exterior side of the protein and exposed to the membrane bilayer; thus, it is unclear if the mutation has any significant impact on the protein's function.[102] The K_a plot aligns with the observed pattern and frequency of amino acids substitutions (Figure 25A). Notably, the K_a plot shows a sudden and steep decrease on December of 2021 corresponding to the Omicron BA.1 variant surge. Starting with the Omicron BA.2 variant surge March 2022, the plot then shows a steep increase and subsequent stability at this new level.

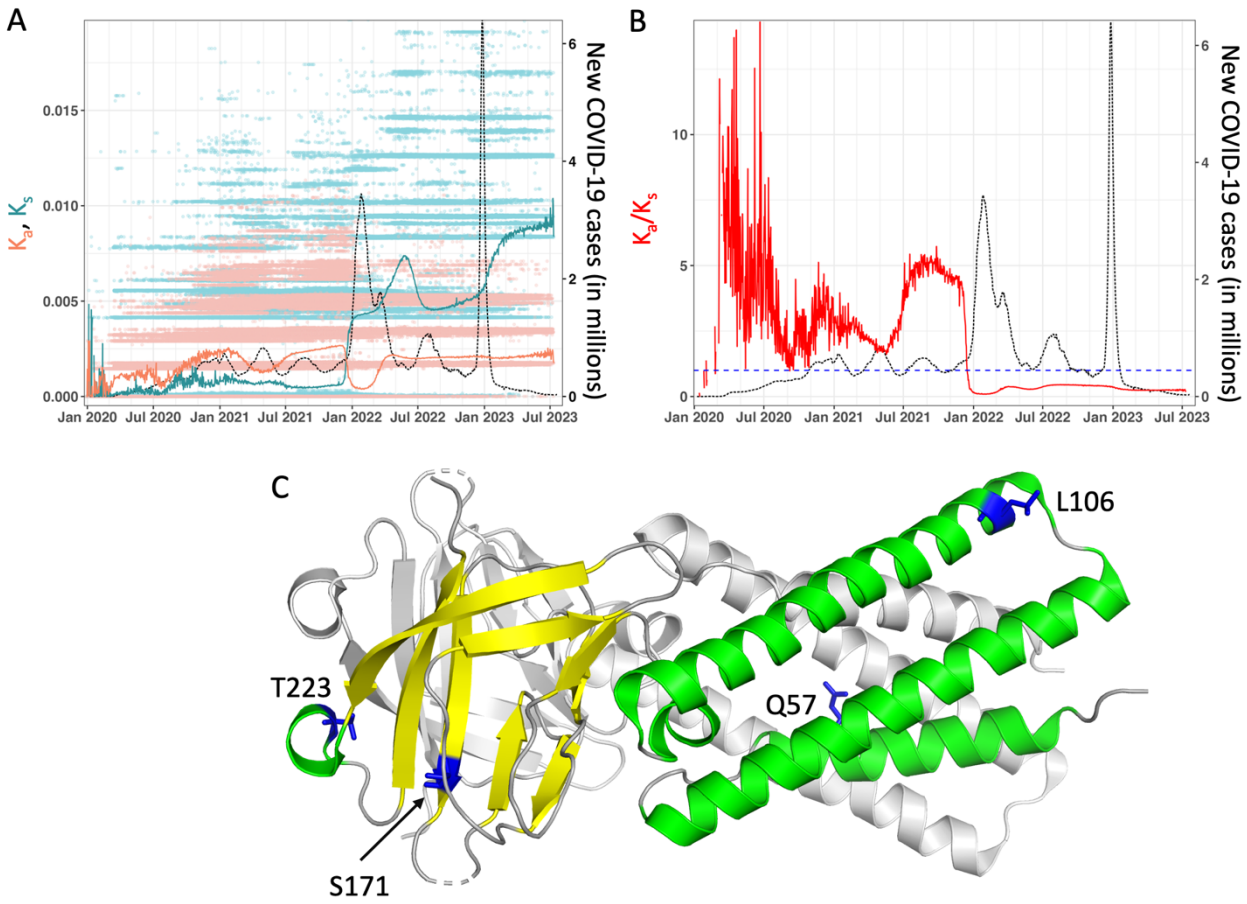

Figure 25: ORF3a mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of ORF3a, based on protein data bank (PDB) ID 6XDC.[103] The structure is shown as a dimer, with only one protomer colored in green and yellow. The missing regions in the structure are indicated by dashed lines.

ORF6. This protein is an antagonist of the interferon-mediated antiviral signaling pathway. Its mode of action is not well understood. However, it was demonstrated that it binds directly to the STAT1 protein resulting in its nuclear exclusion.[104] It has also been observed that ORF6 binds the Nup98 nuclear pore component, resulting in the inhibition of the nuclear translocation of STAT1 and STAT2.[105] More functions are speculated to be attributed to ORF6, as the

localization to different membranes and its inhibitory action to the STATs are independent.[106] We observed a D61L substitution only most Omicron subvariants, except for BA.1, BA.3, and BA.5 (Table 1). No other variants contained mutations. The amino acid substitution pattern observed is mirrored by the K_a plot, as the values increase in March 2022 in conjunction with the rise of the variants containing mutations (Figure 26).

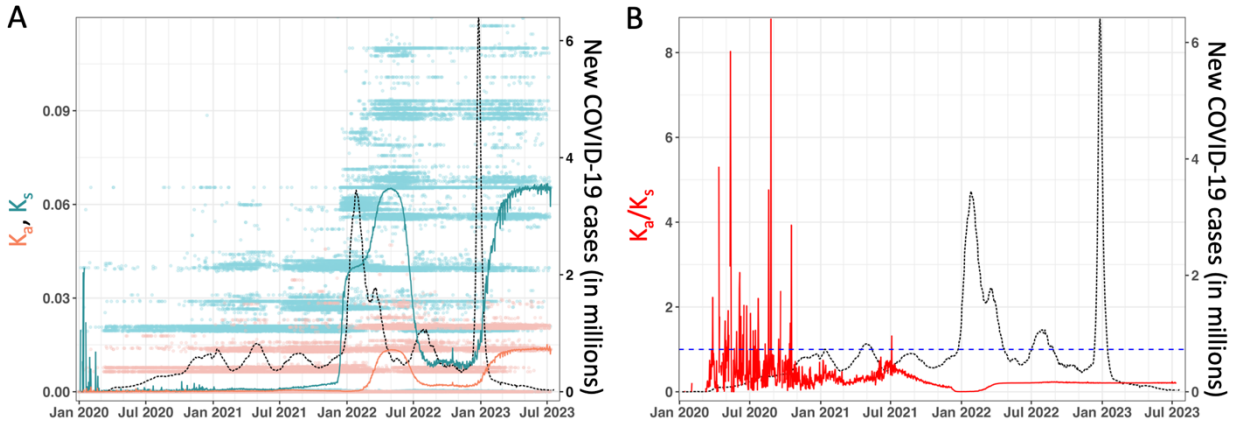

Figure 26: ORF6 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s).

ORF7a. This protein potentially inhibits the immobilization of the host's tethering proteins to allow the mature viral particles to be released outside the host.[107] It was also found to antagonize host interferon-I signaling, primarily by inhibition of STAT2 phosphorylation.[108] In addition, a potential role in disrupting the cell cycle and inducing apoptosis has been found.[19] Recent research has shown that overexpression of ORF7a will result in the accumulation of autophagosomes and prevention of their fusion with the lysosome, which ultimately promotes viral production.[109] Moreover, research suggests the interaction of this protein with ORF3a to evade the immune system.[99] The only variant that has mutations is the Delta variant with two

substitutions: V82A and T120I (Figure 27 and Table 1). The change in amino acid sequence is echoed on the K_a plot, as the graph shows a local increase that spans the Delta variant surge and the decreases again just before the Omicron variant surges.

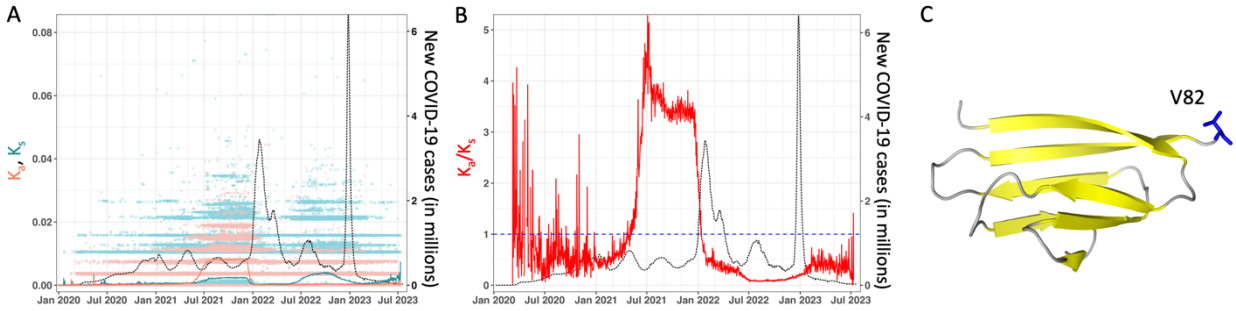

Figure 27: ORF7a mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of ORF7a, based on protein data bank (PDB) ID 7CI3.[110]

ORF7b. This protein is believed to promote apoptosis.[111] Additionally, it can aid in immune system avoidance through antagonism of the interferon-I signaling pathway, through the inhibition of phosphorylation of both STAT1 and STAT2.[108] Surprisingly, the only substitution observed is L11F in the Omicron BA.4 variant (Table 1). The K_a plot reflects the amino acid substitution at the Delta surge (Figure 28). Immediately at the Omicron BA.1 variant surge, we can see the steep decrease of K_a (as well as the K_a/K_s ratio) that remains stable, with only a minute increase around the Omicron BA.4 variant's presence in June 2022.

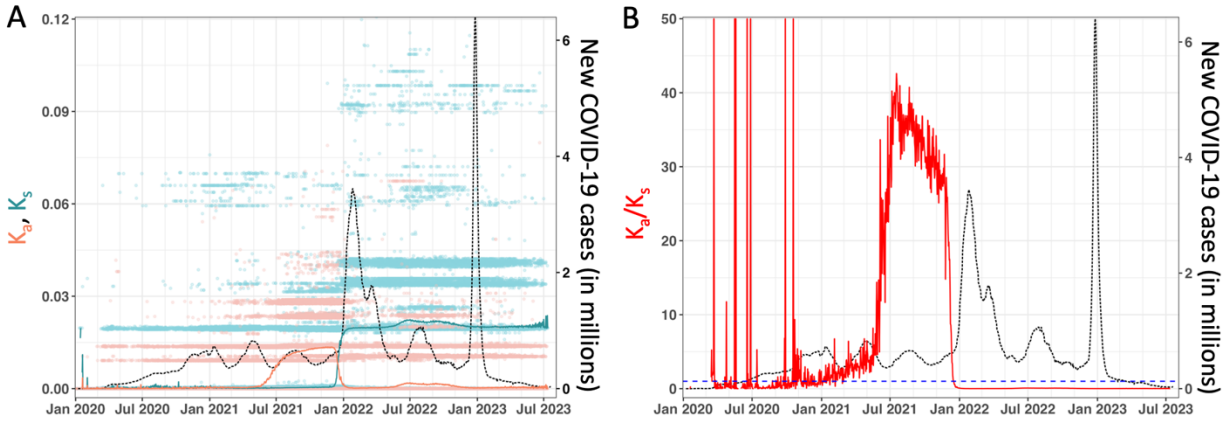

Figure 28: ORF7b mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s).

ORF8. One of the markers for COVID-19 infection is the presence of anti-ORF8 antibodies.[112] ORF8 has been shown to disrupt interferon-I signaling pathway and aid in immune system evasion.[112] Sequence comparison work has found that ORF8 is a paralog of ORF7a, although it is much more dynamic and volatile than the relatively more constrained ORF7a.[113] More recently, ORF8-knockouts have shown a noticeable decrease in lung inflammation in hamsters [108114]. Additionally, that same study demonstrated that a recombinant ORF8 infection would cause lymphocyte infiltration into the lung, causing severe inflammation.[114] The two variants that have mutations pertaining to this protein are the Gamma variant with a E92K non-conservative mutation and the Delta variant with a F120L conservative mutation (Table 1). Of note is that the Alpha variant has a truncated ORF8, which seems to render it non-functional. Mutation values steadily increase until the beginning of the Omicron variant surges in December 2021, where they sharply return to 0 (Figure 29). Interestingly, synonymous mutations sharply outpace nonsynonymous ones immediately prior to the Omicron decrease.

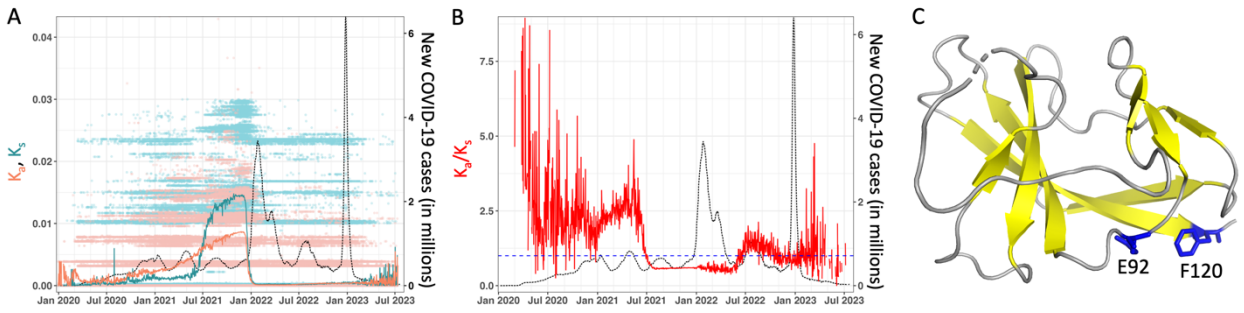

Figure 29: ORF8 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s). (C) Structure of ORF8, based on protein data bank (PDB) ID 7JTL.[115]

ORF10. This is a 38 amino acid-long protein. It's been shown that it binds to E3-ubiquitin ligase, part of the host proteins degradation pathway, where the host proteins ultimately are degraded through the proteasome.[116] This association was found to not be related to the pathogenicity of SARS-CoV-2.[116] However, more recently, ORF10 demonstrated its ability to suppress the interferon-I signaling pathway.[117] Specifically, the mitochondrial antiviral signaling protein (MAVS) was suggested to be the target of ORF10 suppression of the interferon-1 pathway.[117] Our observation of this protein agrees with previous findings suggesting that this gene is under purifying pressure.[118] We noticed only one substitution, A8V, in the Beta variant (Table 1). The paucity of VOC substitutions was reflected in the K_a and K_s graphs (Figure 30). The K_a/K_s ratio graph shows extreme noise in the data, which precludes us from depicting any certain conclusions despite the apparent trend. However, it is possible that mutations were occurring outside of the dominating VOCs.

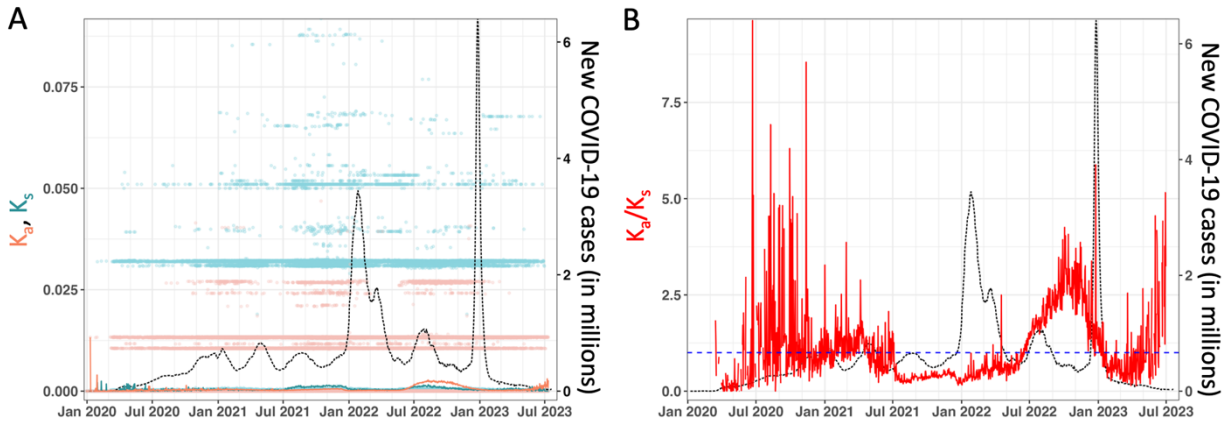

Figure 30: ORF10 mutation summary over the course of the COVID-19 outbreak. (A) Non-synonymous (K_a) and synonymous (K_s) mutations. See Figure 5 for the VOC names corresponding to the peaks in infections and other details. (B) Ratio of non-synonymous to synonymous mutations (K_a/K_s).

CONCLUSIONS AND SUMMARY

Several million samples from the COVID-19 patients have now been sequenced for SARS-CoV-2 genome, and efforts are underway by the wider community to deposit these in public repositories such as NCBI's GenBank and GISAID. As a result, an unprecedented number of SARS-CoV-2 are now available for genomic surveillance. Our group recently reported an approach to detecting mutations changes in real-time and its use for predicting the increase in number of infection cases ahead of time, with the purpose of giving medical community time for pandemic preparedness.

The analysis of over 8 million viral sequences now available in the GenBank indicate that SARS-CoV-2 continues to mutate heavily since the beginning of COVD19 outbreak. The virus is has shown a significant number of mutations, with the largest number of adaptations visible in the spike protein; most noticeably, the number of mutations have increased post-vaccination efforts. Overall, our analysis and results presented here suggest that the mutational propensities

are the largest for the structural proteins, namely the spike, envelope, membrane, and the nucleocapsid proteins. Interestingly, all of these are external proteins that interact with the human immune system. On the other hand, for the majority of the period of the outbreak, the lowest number of mutations are observed in the cofactor proteins that form viral enzyme complexes. More recently, the two proteins NSP1 and NSP13 started showing an increased rate of adaptations, beginning post-Omicron BA period. New mutations have also been observed recently in NSP6, which regulates autophagy. However, NSP7-11 have not shown any significant mutations since the beginning of the pandemic.

From a genomic surveillance for pandemic surge prediction, it appears that ongoing monitoring of K_a provides a fairly reliable metric for surge prediction. It was noticed that for several structural proteins, the rate of mutations increased (and sometimes decreased) 2-4 weeks before the number of reported infection cases. In particular, for a number of surges in infection cases associated with various VOCs there were a number of mutations that occurred in the structural proteins as well as some NSPs. Use of the commonly used K_a/K_s ratio does not appear to provide a reliable signal in most cases, even though it does tend to increase (or decrease) ahead of some surges, but there are also a significant number of cases where increase in this ratio was not followed by a surge. Based on these changes observed so far since the beginning of the pandemic, it appears that the virus continues to evolve, as it is still undergoing a significant number of adaptations. Collectively, based on the observed rates of mutations associated with relevant surges of several VOCs it is possible to make predictions about adaptations that would likely occur in future VOCs (see methods section for details). Table 2 shows a list of mutations that were predicted based on our approach on three dates, where we issued surveillance a watch or warning (full information is available on the website).

Table 2: Prediction of mutations in the upcoming VOC

Jul. 14th 2022	Sep. 7th 2022	Jan. 7th 2023
Spike: T19I, Δ24, Δ25, Δ26, A27S, Δ69, Δ70, G142D, V213G, G339D, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, L452R, S477N, T478K, E484A, F486V, Q498R, N501Y, Y505H, D614G, H655Y, N679K, P681H, N764K, D796Y, Q954H, N969K **Envelope:** T9I **Membrane:** D3N, Q19E, A63T **Nucleocapsid:** P13L, Δ31, Δ32, Δ33, R203K, G204R, S413R **NSP1:** S135R **NSP4:** L264F, T327I, T492I **NSP5:** P132H **NSP6:** Δ105, Δ106, Δ107, F108L **NSP12:** P323L **NSP13:** R392C **NSP14:** I42V **NSP15:** T112I **NSP16:** T140I **ORF10:** L37F	**Spike:** T19I, Δ24, Δ25, Δ26, A27S, Δ69, Δ70, G142D, V213G, G339D, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, L452R, S477N, T478K, E484A, F486V, Q498R, N501Y, Y505H, D614G, H655Y, N679K, P681H, N764K, D796Y, Q954H, N969K **Envelope:** T9I **Membrane:** D3N, Q19E, A63T **NSP1:** S135R **NSP4:** L264F, T327I, T492I **NSP5:** P132H **NSP6:** Δ105, Δ106, Δ107, F108L **NSP12:** P323L **NSP13:** R392C **NSP14:** I42V **NSP15:** T112I **NSP16:** T140I **ORF3a:** T223I **ORF6:** D61L **ORF10:** L37F	**Spike:** L5F, M153T, N164K, H245N, G257D, K444R, N450D, L452M, N460K, E484R, G485D **NSP2:** T547I **NSP3:** I896T, G1273D, M1351V, T1356I, T1830I **NSP4:** A146V, L438F **NSP7:** Q63R **NSP12:** D155G **NSP13:** T481M **ORF3a:** L140F

Mutations predicted based on method described in Figure 2. See https://pandemics.okstate.edu/covid19/ for more details.

The presented framework and the real-time project website enable the tracking of the new mutations in real-time and possibly enable prediction of future variants which are relevant for diagnostic kits, as large changes could potentially make detection by the antibodies difficult. Similarly, for the vaccine and drug design efforts it would be important to keep an eye on the changes. The research community gained a wealth of knowledge about SARS-CoV-2 and its genome due to a considerable effort world-wide to tackle the pandemic, with sustained efforts to sequence millions of sequences enabling genomic surveillance efforts. Arguably, perhaps the community learned more about the genome of this virus in a shorter period of time than any other virus because of efforts across the globe to both sequence and deposit results in public databases. The recent downturn in the number of samples from COVID-19 positive patients being sequenced and being publicly being reported could undermine the genomic surveillance efforts. Therefore, the presented results serve as a call to the medical and health community to keep dedicating resources for regularly sequencing the virus. This prototype would also serve as

a model for future pandemics, where the presented framework could easily be reworked and immediately be adapted from day one.

References

1 Finkel, Y. *et al.* The coding capacity of SARS-CoV-2. *Nature* **589**, 125-130, doi:10.1038/s41586-020-2739-1 (2021).

2 da Silva, S. J. R., Alves da Silva, C. T., Mendes, R. P. G. & Pena, L. Role of nonstructural proteins in the pathogenesis of SARS-CoV-2. *J Med Virol* **92**, 1427-1429, doi:10.1002/jmv.25858 (2020).

3 Davidson, A. D. *et al.* Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. *Genome Med* **12**, 68, doi:10.1186/s13073-020-00763-0 (2020).

4 Mathieu, E. *et al.* Coronavirus Pandemic (COVID-19). *OurWorldInData.org* (2020).

5 (US), N. C. f. B. I. *Entrez Programming Utilities Help*. (2010).

6 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, doi:10.1038/msb.2011.75 (2011).

7 Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609-612, doi:10.1093/nar/gkl315 (2006).

8 Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**, 77-80, doi:10.1016/S1672-0229(10)60008-3 (2010).

9 Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431-449, doi:10.1093/genetics/155.1.431 (2000).

10 Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**, 496-503, doi:10.1016/s0169-5347(00)01994-7 (2000).

11 Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028, doi:10.1038/nbt.3988 (2017).

12 Xia, X. Domains and Functions of Spike Protein in Sars-Cov-2 in the Context of Vaccine Design. *Viruses* **13**, doi:10.3390/v13010109 (2021).

13 Pillay, T. S. Gene of the month: the 2019-nCoV/SARS-CoV-2 novel coronavirus spike protein. *J Clin Pathol* **73**, 366-369, doi:10.1136/jclinpath-2020-206658 (2020).

14 Wang, Q. *et al.* Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* **181**, 894-904 e899, doi:10.1016/j.cell.2020.03.045 (2020).

15 Onnis, A. *et al.* SARS-CoV-2 Spike protein suppresses CTL-mediated killing by inhibiting immune synapse assembly. *J Exp Med* **220**, doi:10.1084/jem.20220906 (2023).

16 Najar, F. Z. *et al.* Future COVID19 surges prediction based on SARS-CoV-2 mutations surveillance. *Elife* **12**, doi:10.7554/eLife.82980 (2023).

17 Zhang, S. *et al.* Bat and pangolin coronavirus spike glycoprotein structures provide insights into SARS-CoV-2 evolution. *Nat Commun* **12**, 1607, doi:10.1038/s41467-021-21767-3 (2021).

18 Zheng, M. *et al.* TLR2 senses the SARS-CoV-2 envelope protein to produce inflammatory cytokines. *Nat Immunol* **22**, 829-838, doi:10.1038/s41590-021-00937-x (2021).

19 Wong, N. A. & Saier, M. H., Jr. The SARS-Coronavirus Infection Cycle: A Survey of Viral Membrane Proteins, Their Functional Interactions and Pathogenesis. *Int J Mol Sci* **22**, doi:10.3390/ijms22031308 (2021).

20 Planes, R., Bert, J. B., Tairi, S., BenMohamed, L. & Bahraoui, E. SARS-CoV-2 Envelope (E) Protein Binds and Activates TLR2 Pathway: A Novel Molecular Target for COVID-19 Interventions. *Viruses* **14**, doi:10.3390/v14050999 (2022).

21 Waisner, H. *et al.* SARS-CoV-2 Harnesses Host Translational Shutoff and Autophagy To Optimize Virus Yields: the Role of the Envelope (E) Protein. *Microbiol Spectr*, e0370722, doi:10.1128/spectrum.03707-22 (2023).

22 Rahman, M. S. *et al.* Mutational insights into the envelope protein of SARS-CoV-2. *Gene Rep* **22**, 100997, doi:10.1016/j.genrep.2020.100997 (2021).

23 Xia, B. *et al.* Why is the SARS-CoV-2 Omicron variant milder? *Innovation (Camb)* **3**, 100251, doi:10.1016/j.xinn.2022.100251 (2022).

24 Li, Y., Surya, W., Claudine, S. & Torres, J. Structure of a conserved Golgi complex-targeting signal in coronavirus envelope proteins. *J Biol Chem* **289**, 12535-12549, doi:10.1074/jbc.M114.560094 (2014).

25 Lu, S. *et al.* The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat Commun* **12**, 502, doi:10.1038/s41467-020-20768-y (2021).

26 Marques-Pereira, C. *et al.* SARS-CoV-2 membrane protein: from genomic data to structural new insights. *Research Square* **PREPRINT**, doi:10.21203/rs.3.rs-702792/v2 (2022).

27 Shen, L. *et al.* Emerging variants of concern in SARS-CoV-2 membrane protein: a highly conserved target with potential pathological and therapeutic implications. *Emerg Microbes Infect* **10**, 885-893, doi:10.1080/22221751.2021.1922097 (2021).

28 Dolan, K. A. *et al.* Structure of SARS-CoV-2 M protein in lipid nanodiscs. *Elife* **11**, doi:10.7554/eLife.81702 (2022).

29 Ye, Q., West, A. M. V., Silletti, S. & Corbett, K. D. Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. *Protein Sci* **29**, 1890-1901, doi:10.1002/pro.3909 (2020).

30 Khan, M. T. *et al.* Structures of SARS-CoV-2 RNA-Binding Proteins and Therapeutic Targets. *Intervirology* **64**, 55-68, doi:10.1159/000513686 (2021).

31 Khan, M. T. *et al.* SARS-CoV-2 nucleocapsid and Nsp3 binding: an in silico study. *Arch Microbiol* **203**, 59-66, doi:10.1007/s00203-020-01998-6 (2021).

32 Rak, A. *et al.* Assessment of Immunogenic and Antigenic Properties of Recombinant Nucleocapsid Proteins of Five SARS-CoV-2 Variants in a Mouse Model. *Viruses* **15**, doi:10.3390/v15010230 (2023).

33 Johnson, B. A. *et al.* Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. *bioRxiv*, doi:10.1101/2021.10.14.464390 (2022).

34 Zhao, H. *et al.* Plasticity in structure and assembly of SARS-CoV-2 nucleocapsid protein. *bioRxiv*, doi:10.1101/2022.02.08.479556 (2022).

35 Bessieres, M. H. *et al.* [Local production of specific antibodies in the aqueous humor in experimental Candida endophthalmia in rabbits]. *Ann Biol Clin (Paris)* **45**, 651-656 (1987).

36 Rafael Ciges-Tomas, J., Franco, M. L. & Vilar, M. Identification of a guanine-specific pocket in the protein N of SARS-CoV-2. *Commun Biol* **5**, 711, doi:10.1038/s42003-022-03647-8 (2022).

37 Bujanic, L. *et al.* The key features of SARS-CoV-2 leader and NSP1 required for viral escape of NSP1-mediated repression. *RNA* **28**, 766-779, doi:10.1261/rna.079086.121 (2022).

38 Schubert, K. *et al.* SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation. *Nat Struct Mol Biol* **27**, 959-966, doi:10.1038/s41594-020-0511-8 (2020).

39 Thoms, M. *et al.* Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. *Science* **369**, 1249-1255, doi:10.1126/science.abc8665 (2020).

40 Wang, Y., Kirkpatrick, J., Lage, S. Z. & Carlomagno, T. Structural insights into the activity regulation of full-length non-structural protein 1 from SARS-CoV-2. *Structure* **31**, 128-137 e125, doi:10.1016/j.str.2022.12.006 (2023).

41 Simeoni, M., Cavinato, T., Rodriguez, D. & Gatfield, D. I(nsp1)ecting SARS-CoV-2-ribosome interactions. *Commun Biol* **4**, 715, doi:10.1038/s42003-021-02265-0 (2021).

42 Li, Y. *et al.* SARS-CoV-2-encoded inhibitors of human LINE-1 retrotransposition. *J Med Virol* **95**, e28135, doi:10.1002/jmv.28135 (2023).

43 Semper, C., Watanabe, N. & Savchenko, A. Structural characterization of nonstructural protein 1 from SARS-CoV-2. *iScience* **24**, 101903, doi:10.1016/j.isci.2020.101903 (2021).

44 Gupta, M. *et al.* CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes. *bioRxiv*, doi:10.1101/2021.05.10.443524 (2021).

45 Shi, F. S. *et al.* Expression Profile and Localization of SARS-CoV-2 Nonstructural Replicase Proteins in Infected Cells. *Microbiol Spectr*, e0074422, doi:10.1128/spectrum.00744-22 (2022).

46 Zou, L., Moch, C., Graille, M. & Chapat, C. The SARS-CoV-2 protein NSP2 impairs the silencing capacity of the human 4EHP-GIGYF2 complex. *iScience* **25**, 104646, doi:10.1016/j.isci.2022.104646 (2022).

47 Periwal, N. *et al.* Time Series Analysis of SARS-CoV-2 Genomes and Correlations among Highly Prevalent Mutations. *Microbiol Spectr* **10**, e0121922, doi:10.1128/spectrum.01219-22 (2022).

48 Ji, M., Li, M., Sun, L., Deng, H. & Zhao, Y. G. DMV biogenesis during beta-coronavirus infection requires autophagy proteins VMP1 and TMEM41B. *Autophagy* **19**, 737-738, doi:10.1080/15548627.2022.2103783 (2023).

49 Rut, W. *et al.* Activity profiling and crystal structures of inhibitor-bound SARS-CoV-2 papain-like protease: A framework for anti-COVID-19 drug design. *Sci Adv* **6**, doi:10.1126/sciadv.abd4596 (2020).

50 Gahbauer, S. *et al.* Iterative computational design and crystallographic screening identifies potent inhibitors targeting the Nsp3 macrodomain of SARS-CoV-2. *Proc Natl Acad Sci U S A* **120**, e2212931120, doi:10.1073/pnas.2212931120 (2023).

51 Gao, X. *et al.* Crystal structure of SARS-CoV-2 papain-like protease. *Acta Pharm Sin B* **11**, 237-245, doi:10.1016/j.apsb.2020.08.014 (2021).

52 Arnaud, M. J. *et al.* Synthesis, effectiveness and metabolic fate in cows of the caesium complexing compound ammonium ferric hexacyanoferrate labelled with 14C. *J Dairy Res* **55**, 1-13, doi:10.1017/s0022029900025796 (1988).

53 Tan, J. *et al.* The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. *PLoS Pathog* **5**, e1000428, doi:10.1371/journal.ppat.1000428 (2009).

54 Berrio, A., Gartner, V. & Wray, G. A. Positive selection within the genomes of SARS-CoV-2 and other Coronaviruses independent of impact on protein function. *PeerJ* **8**, e10234, doi:10.7717/peerj.10234 (2020).

55 Xu, X. *et al.* Crystal structure of the C-terminal cytoplasmic domain of non-structural protein 4 from mouse hepatitis virus A59. *PLoS One* **4**, e6217, doi:10.1371/journal.pone.0006217 (2009).

56 Angelini, M. M., Akhlaghpour, M., Neuman, B. W. & Buchmeier, M. J. Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. *mBio* **4**, doi:10.1128/mBio.00524-13 (2013).

57 fooladinezhad, H. *et al.* SARS-CoV-2 NSP3, NSP4 and NSP6 mutations and Epistasis during the pandemic in the world: Evolutionary Trends and Natural Selections in Six Continents. *medRxiv*, 2022.2005.2022.22275422, doi:10.1101/2022.05.22.22275422 (2022).

58 Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors. *Science* **368**, 409-412, doi:10.1126/science.abb3405 (2020).

59 Chen, J. *et al.* SARS-CoV-2 nsp5 Exhibits Stronger Catalytic Activity and Interferon Antagonism than Its SARS-CoV Ortholog. *J Virol* **96**, e0003722, doi:10.1128/jvi.00037-22 (2022).

60 Ullrich, S., Ekanayake, K. B., Otting, G. & Nitsche, C. Main protease mutants of SARS-CoV-2 variants remain susceptible to nirmatrelvir. *Bioorg Med Chem Lett* **62**, 128629, doi:10.1016/j.bmcl.2022.128629 (2022).

61 Sacco, M. D. *et al.* The P132H mutation in the main protease of Omicron SARS-CoV-2 decreases thermal stability without compromising catalysis or small-molecule drug inhibition. *Cell Res* **32**, 498-500, doi:10.1038/s41422-022-00640-y (2022).

62 Xia, Z. *et al.* Rational Design of Hybrid SARS-CoV-2 Main Protease Inhibitors Guided by the Superimposed Cocrystal Structures with the Peptidomimetic Inhibitors GC-376, Telaprevir, and Boceprevir. *ACS Pharmacol Transl Sci* **4**, 1408-1421, doi:10.1021/acsptsci.1c00099 (2021).

63 Cottam, E. M., Whelband, M. C. & Wileman, T. Coronavirus NSP6 restricts autophagosome expansion. *Autophagy* **10**, 1426-1441, doi:10.4161/auto.29309 (2014).

64 Wang, R., Chen, J., Hozumi, Y., Yin, C. & Wei, G. W. Decoding Asymptomatic COVID-19 Infection and Transmission. *J Phys Chem Lett* **11**, 10007-10015, doi:10.1021/acs.jpclett.0c02765 (2020).

65 Chen, D. Y. *et al.* Spike and nsp6 are key determinants of SARS-CoV-2 Omicron BA.1 attenuation. *Nature*, doi:10.1038/s41586-023-05697-2 (2023).

66 Khalid, M., Murphy, D., Shoai, M., George-William, J. N. & Al-Ebini, Y. Geographical distribution of host's specific SARS-CoV-2 mutations in the early phase of the COVID-19 pandemic. *Gene* **851**, 147020, doi:10.1016/j.gene.2022.147020 (2023).

67 te Velthuis, A. J., Arnold, J. J., Cameron, C. E., van den Worm, S. H. & Snijder, E. J. The RNA polymerase activity of SARS-coronavirus nsp12 is primer dependent. *Nucleic Acids Res* **38**, 203-214, doi:10.1093/nar/gkp904 (2010).

68 Peng, Q. *et al.* Structural and Biochemical Characterization of the nsp12-nsp7-nsp8 Core Polymerase Complex from SARS-CoV-2. *Cell Rep* **31**, 107774, doi:10.1016/j.celrep.2020.107774 (2020).

69 Biswal, M. *et al.* Two conserved oligomer interfaces of NSP7 and NSP8 underpin the dynamic assembly of SARS-CoV-2 RdRP. *Nucleic Acids Res* **49**, 5956-5966, doi:10.1093/nar/gkab370 (2021).

70 Makiyama, K. *et al.* NSP9 of SARS-CoV-2 attenuates nuclear transport by hampering nucleoporin 62 dynamics and functions in host cells. *Biochem Biophys Res Commun* **586**, 137-142, doi:10.1016/j.bbrc.2021.11.046 (2022).

71 Banerjee, A. K. *et al.* SARS-CoV-2 Disrupts Splicing, Translation, and Protein Trafficking to Suppress Host Defenses. *Cell* **183**, 1325-1339 e1321, doi:10.1016/j.cell.2020.10.004 (2020).

72 Chasapis, C. T., Perlepes, S. P., Bjorklund, G. & Peana, M. Structural modeling of protein ensembles between E3 RING ligases and SARS-CoV-2: The role of zinc binding domains. *J Trace Elem Med Biol* **75**, 127089, doi:10.1016/j.jtemb.2022.127089 (2023).

73 Zhang, C. *et al.* Structural basis for the multimerization of nonstructural protein nsp9 from SARS-CoV-2. *Mol Biomed* **1**, 5, doi:10.1186/s43556-020-00005-0 (2020).

74 Riccio, A. A., Sullivan, E. D. & Copeland, W. C. Activation of the SARS-CoV-2 NSP14 3'-5' exoribonuclease by NSP10 and response to antiviral inhibitors. *J Biol Chem* **298**, 101518, doi:10.1016/j.jbc.2021.101518 (2022).

75 Arabi-Jeshvaghani, F., Javadi-Zarnaghi, F. & Ganjalikhany, M. R. Analysis of critical protein-protein interactions of SARS-CoV-2 capping and proofreading molecular machineries towards designing dual target inhibitory peptides. *Sci Rep* **13**, 350, doi:10.1038/s41598-022-26778-8 (2023).

76 Hamre, J. R., 3rd & Jafri, M. S. Optimizing peptide inhibitors of SARS-Cov-2 nsp10/nsp16 methyltransferase predicted through molecular simulation and machine learning. *Inform Med Unlocked* **29**, 100886, doi:10.1016/j.imu.2022.100886 (2022).

77 Moeller, N. H. *et al.* Structure and dynamics of SARS-CoV-2 proofreading exoribonuclease ExoN. *Proc Natl Acad Sci U S A* **119**, doi:10.1073/pnas.2106379119 (2022).

78 Gadhave, K. *et al.* Conformational dynamics of 13 amino acids long NSP11 of SARS-CoV-2 under membrane mimetics and different solvent conditions. *Microb Pathog* **158**, 105041, doi:10.1016/j.micpath.2021.105041 (2021).

79 Petushkov, I., Esyunina, D. & Kulbachinskiy, A. Effects of natural RNA modifications on the activity of SARS-CoV-2 RNA-dependent RNA polymerase. *FEBS J* **290**, 80-92, doi:10.1111/febs.16587 (2023).

80 Alam, A. *et al.* Dominant clade-featured SARS-CoV-2 co-occurring mutations reveal plausible epistasis: An in silico based hypothetical model. *J Med Virol* **94**, 1035-1049, doi:10.1002/jmv.27416 (2022).

81 Malone, B. *et al.* Structural basis for backtracking by the SARS-CoV-2 replication-transcription complex. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2102516118 (2021).

82 Fung, S. Y. *et al.* SARS-CoV-2 NSP13 helicase suppresses interferon signaling by perturbing JAK1 phosphorylation of STAT1. *Cell Biosci* **12**, 36, doi:10.1186/s13578-022-00770-1 (2022).

83 Rashid, F. *et al.* Structural Analysis on the Severe Acute Respiratory Syndrome Coronavirus 2 Non-structural Protein 13 Mutants Revealed Altered Bonding Network With TANK Binding Kinase 1 to Evade Host Immune System. *Front Microbiol* **12**, 789062, doi:10.3389/fmicb.2021.789062 (2021).

84 Kumari, D. *et al.* Identification and Characterization of Novel Mutants of Nsp13 Protein among Indian SARS-CoV-2 Isolates. *The Open Bioinformatics Journal* **15**, doi:10.2174/18750362-v15-e2202100 (2022).

85 Newman, J. A. *et al.* Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nat Commun* **12**, 4848, doi:10.1038/s41467-021-25166-6 (2021).

86 Ahmed-Belkacem, R. *et al.* Potent Inhibition of SARS-CoV-2 nsp14 N7-Methyltransferase by Sulfonamide-Based Bisubstrate Analogues. *J Med Chem* **65**, 6231-6249, doi:10.1021/acs.jmedchem.2c00120 (2022).

87 Yan, L. *et al.* Coupling of N7-methyltransferase and 3'-5' exoribonuclease with SARS-CoV-2 polymerase reveals mechanisms for capping and proofreading. *Cell* **184**, 3474-3485 e3411, doi:10.1016/j.cell.2021.05.033 (2021).

88 Yuen, C. K. *et al.* SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 function as potent interferon antagonists. *Emerg Microbes Infect* **9**, 1418-1428, doi:10.1080/22221751.2020.1780953 (2020).

89 Liu, C. *et al.* Structural basis of mismatch recognition by a SARS-CoV-2 proofreading enzyme. *Science* **373**, 1142-1146, doi:10.1126/science.abi9310 (2021).

90 Bhardwaj, K., Sun, J., Holzenburg, A., Guarino, L. A. & Kao, C. C. RNA recognition and cleavage by the SARS coronavirus endoribonuclease. *J Mol Biol* **361**, 243-256, doi:10.1016/j.jmb.2006.06.021 (2006).

91 Ivanov, K. A. *et al.* Major genetic marker of nidoviruses encodes a replicative endoribonuclease. *Proc Natl Acad Sci U S A* **101**, 12694-12699, doi:10.1073/pnas.0403127101 (2004).

92 Saramago, M. *et al.* The nsp15 Nuclease as a Good Target to Combat SARS-CoV-2: Mechanism of Action and Its Inactivation with FDA-Approved Drugs. *Microorganisms* **10**, doi:10.3390/microorganisms10020342 (2022).

93 Deng, X. *et al.* Coronavirus nonstructural protein 15 mediates evasion of dsRNA sensors and limits apoptosis in macrophages. *Proc Natl Acad Sci U S A* **114**, E4251-E4260, doi:10.1073/pnas.1618310114 (2017).

94 Deng, X. & Baker, S. C. An "Old" protein with a new story: Coronavirus endoribonuclease is important for evading host antiviral defenses. *Virology* **517**, 157-163, doi:10.1016/j.virol.2017.12.024 (2018).

95 Vithani, N. *et al.* SARS-CoV-2 Nsp16 activation mechanism and a cryptic pocket with pan-coronavirus antiviral potential. *Biophys J* **120**, 2880-2889, doi:10.1016/j.bpj.2021.03.024 (2021).

96 Viswanathan, T. *et al.* Structural basis of RNA cap modification by SARS-CoV-2. *Nat Commun* **11**, 3718, doi:10.1038/s41467-020-17496-8 (2020).

97 Bianchi, M., Borsetti, A., Ciccozzi, M. & Pascarella, S. SARS-Cov-2 ORF3a: Mutability and function. *Int J Biol Macromol* **170**, 820-826, doi:10.1016/j.ijbiomac.2020.12.142 (2021).

98 Chen, D. *et al.* ORF3a of SARS-CoV-2 promotes lysosomal exocytosis-mediated viral egress. *Dev Cell* **56**, 3250-3263 e3255, doi:10.1016/j.devcel.2021.10.006 (2021).

99 Arshad, N. *et al.* SARS-CoV-2 accessory proteins ORF7a and ORF3a use distinct mechanisms to down-regulate MHC-I surface expression. *Proc Natl Acad Sci U S A* **120**, e2208525120, doi:10.1073/pnas.2208525120 (2023).

100 Nagy, A., Pongor, S. & Gyorffy, B. Different mutations in SARS-CoV-2 associate with severe and mild outcome. *Int J Antimicrob Agents* **57**, 106272, doi:10.1016/j.ijantimicag.2020.106272 (2021).

101 Zhang, J. *et al.* Genome-Wide Characterization of SARS-CoV-2 Cytopathogenic Proteins in the Search of Antiviral Targets. *mBio* **13**, e0016922, doi:10.1128/mbio.00169-22 (2022).

102 Kern, D. M. *et al.* Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs. *Nat Struct Mol Biol* **28**, 573-582, doi:10.1038/s41594-021-00619-0 (2021).

103 Kern, D. M. *et al.* Cryo-EM structure of the SARS-CoV-2 3a ion channel in lipid nanodiscs. *bioRxiv*, doi:10.1101/2020.06.17.156554 (2021).

104 Miyamoto, Y. *et al.* SARS-CoV-2 ORF6 disrupts nucleocytoplasmic trafficking to advance viral replication. *Commun Biol* **5**, 483, doi:10.1038/s42003-022-03427-4 (2022).

105 Miorin, L. *et al.* SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling. *Proc Natl Acad Sci U S A* **117**, 28344-28354, doi:10.1073/pnas.2016650117 (2020).

106 Wong, H. T., Cheung, V. & Salamango, D. J. Decoupling SARS-CoV-2 ORF6 localization and interferon antagonism. *J Cell Sci* **135**, doi:10.1242/jcs.259666 (2022).

107 Petrosino, M. *et al.* Zn-Induced Interactions Between SARS-CoV-2 orf7a and BST2/Tetherin. *ChemistryOpen* **10**, 1133-1141, doi:10.1002/open.202100217 (2021).

108 Xia, H. *et al.* Evasion of Type I Interferon by SARS-CoV-2. *Cell Rep* **33**, 108234, doi:10.1016/j.celrep.2020.108234 (2020).

109 Hou, P. *et al.* The ORF7a protein of SARS-CoV-2 initiates autophagy and limits autophagosome-lysosome fusion via degradation of SNAP29 to promote virus replication. *Autophagy* **19**, 551-569, doi:10.1080/15548627.2022.2084686 (2023).

110 Zhou, Z. *et al.* Structural insight reveals SARS-CoV-2 ORF7a as an immunomodulating factor for human CD14(+) monocytes. *iScience* **24**, 102187, doi:10.1016/j.isci.2021.102187 (2021).

111 Yang, R. *et al.* SARS-CoV-2 Accessory Protein ORF7b Mediates Tumor Necrosis Factor-alpha-Induced Apoptosis in Cells. *Front Microbiol* **12**, 654709, doi:10.3389/fmicb.2021.654709 (2021).

112 Flower, T. G. *et al.* Structure of SARS-CoV-2 ORF8, a rapidly evolving immune evasion protein. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2021785118 (2021).

113 Neches, R. Y., Kyrpides, N. C. & Ouzounis, C. A. Atypical Divergence of SARS-CoV-2 Orf8 from Orf7a within the Coronavirus Lineage Suggests Potential Stealthy Viral Strategies in Immune Evasion. *mBio* **12**, doi:10.1128/mBio.03014-20 (2021).

114 Kohyama, M. *et al.* SARS-CoV-2 ORF8 is a viral cytokine regulating immune responses. *Int Immunol* **35**, 43-52, doi:10.1093/intimm/dxac044 (2023).

115 Leads from the MMWR. Regional differences in postneonatal mortality--Mississippi, 1980-1983. *JAMA* **259**, 186, 189 (1988).

116 Mena, E. L. *et al.* ORF10-Cullin-2-ZYG11B complex is not required for SARS-CoV-2 infection. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2023157118 (2021).

117 Li, X. *et al.* SARS-CoV-2 ORF10 suppresses the antiviral innate immune response by degrading MAVS through mitophagy. *Cell Mol Immunol* **19**, 67-78, doi:10.1038/s41423-021-00807-4 (2022).

118 Cagliani, R., Forni, D., Clerici, M. & Sironi, M. Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses. *Infect Genet Evol* **83**, 104353, doi:10.1016/j.meegid.2020.104353 (2020).