Future COVID19 surges prediction based on SARS-CoV-2 mutations surveillance

*Fares Z. Najar,[1] Evan Linde,[1] Chelsea L. Murphy,[1] Veniamin A. Borin,[1,2] 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

*Corresponding author
pratul.agarwal@okstate.edu, 405 744-6639
MS 106, Oklahoma State University, Stillwater OK 74078

ABSTRACT

COVID19 has aptly revealed that airborne viruses such as SARS-CoV-2 with the ability to rapidly mutate, combined with high rates of transmission and fatality can cause a deadly world-wide pandemic in a matter of weeks.[1] Apart from vaccines and post-infection treatment options, strategies for preparedness will be vital in responding to the current and future pandemics. Therefore, there is wide interest in approaches that allow predictions of increase in infections ("surges") before they occur. We describe here real time genomic surveillance particularly based on mutation analysis, of viral proteins as a methodology for *a priori* determination of surge in number of infection cases. The full results are available for SARS-CoV-2 at http://pandemics.okstate.edu/covid19/, and are updated daily as new virus sequences become available. This approach is generic and will also be applicable to other pathogens.

INTRODUCTION

Protein and genome sequence analyses identify molecular level changes that enable viral adaptations for increased spread through the host population. Concrete evidence for direct a relationship between specific mutations and increase in rates of infection (and fatality) requires extensive laboratory studies that need significant time. The availability of unprecedented number of SARS-CoV-2 genome sequences is making possible identification of number and types of mutations, which in turn can provide vital knowledge in real time, crucial for decision making by health professionals for medical interventions. We are investigating several different approaches (synonymous, non-synonymous, and non-synonymous/ synonymous ratio for the nucleotide sequences,[2] and conservative or radical substitutions for the amino acid sequences) for using number and types of mutations as a means to predict surge in infections as well as monitor the changes in critical viral proteins. Recently, such analysis has been reported for single SARS-CoV-2 proteins.[3] Our approach, however, is based on the whole viral genome analysis and moreover it is performed continually in real time.

MATERIALS AND METHODS

The SARS-CoV-2 genomic sequences data and the number of COVID19 sequences are continually obtained from the sources described below. The genomic sequences are carefully filtered for quality control and used for calculations of non-synonymous (k_a) and synonymous (k_s) mutation rates for each of the 26 proteins separately.

Data and data sources: Data for the number of reported COVID19 cases was accessed from Johns Hopkins University's 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. 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 6,468,196 records were searched and a total of 3,126,129 sequences matching the search criterion and passing the quality control steps were used as of November 21th, 2022.

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 version 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 with the nucleotide sequences, the associated protein sequences, the collection dates, and the country of collection were stored in SQLite relational database where they were indexed with unique identifiers to allow the retrieval and analysis of any part of the parsed data.

Frequency of data updates: As of July 2022, the described sources are monitored daily for updates. New data is continually downloaded and used for analysis.

Alignments and non-synonymous (k_a), synonymous (k_s) calculations: The translated proteins and nucleotides 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 and non-synonymous (k_a), and synonymous (k_s), 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* 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")[10] was used as a reference for all the k_a, k_s and k_a/k_s calculations. We explored the possibility of using other sequence(s) as references (for example, the previous day or the previous month), however, due to the increasing number of variations available every day, it is difficult to select a representative sequence on an ongoing basis. It was also found that using the Wuhan sequence as a reference provided the most intuitive and interpretable results.

List of proteins investigated: The number of unique nucleotide sequences observed till date for each of the 26 proteins/open reading frames are listed in Table 1 below. The full results are available on the project website https://pandemics.okstate.edu/covid19/, which are continually updated.

Table 1: Number of unique records for the 26 proteins/open reading frames. Total number of quality-controlled SARS-CoV-2 sequences analyzed: 3,126,129 (as of November 21th, 2022). Only three proteins showing the most relevant results and one protein (marked by *) for comparison is depicted in Figures. These proteins are shown in bold.

Name	Unique records
Envelope protein	**1,314**
Membrane protein	**11,338**
Nucleocapsid protein	70,579
Spike protein	**188,166**
Non-structural protein 1 (NSP1), leader protein	11,656
NSP2	67,837
NSP3	245,627
NSP4	31,257
NSP5, 3C-like Proteinase	11,879
NSP6	16,479
NSP7	1,304
NSP8	4,490
NSP9	2,848
NSP10	2,429
NSP11	88
NSP12, RNA-dependent RNA polymerase (RDRP)*	**60,575**
NSP13, helicase	35,421
NSP14, 3'-to-5' exonuclease	28,501
NSP15, endoRNAse	12,901
NSP16, 2'-O-ribose methyltransferase	7,636
ORF3a	41,694
ORF6	2,117
ORF7a	9,312
ORF7b	1,368
ORF8	7.036
ORF10	710

RESULTS

It was found that collective non-synonymous mutations in key proteins of SARS-CoV-2 showed significant increase 10-14 days before the rapid rise in COVID19 cases, particularly related to the surges that occurred after the emergence of Gamma, Delta, Omicron and BA5 variants (Figure 1 and the related Figure 1-figure supplement 1 with the unnormalized results). At present, over 6.4 million SARS-CoV-2 genome sequences collected all over the world are available from GenBank (https://www.ncbi.nlm.nih.gov/sars-cov-2/), which were used for analysis of 26 SARS-CoV-2 proteins including the structural (spike, envelope, membrane, nucleocapsid) proteins, non-structural proteins (NSPs) and open reading frames (ORFs). Note, our analysis was performed with the first reported ("Wuhan") SARS-CoV-2 sequence as a reference.[10] In other words, the computed mutations are calculated in comparison to this reference sequence. The reason for an increase in mutations ahead of a surge is the search for adaptation against the acquired immunity (or gain in function) in either a single protein or a combination of proteins. The case of the Omicron variant indicates the development of the most drastic changes in several different proteins, which coincided with the largest increase in rate of infections (Figure 1). Non-synonymous mutations (k_a) in several proteins show significant increase before the increase in rate of infections (or surges), therefore, allowing a means for surge prediction.

Use of mutational rates as a surge predictor: In addition to using non-synonymous mutations, a number of other metrics were also investigated for a reliable prediction signal. In particular, the commonly used non-synonymous to synonymous mutations ratio, k_a/k_s, (Figure 1-figure supplement 2) and the rate of mutations (derivative of observed number of mutations with respect to time) (Figure 1-figure supplement 3) were also investigated in detail for suitability as a signal for surge prediction. As shown in Figure 1-figure supplement 2, k_a/k_s did not provide a reliable surge prediction signal. Figure 1-figure supplement 3 shows rate of mutations (calculated as a numerical derivative). For the case of Omicron surge, the proteins did show increased rate of mutations, however, for all other cases a clear signal was absent. Furthermore, the rate of mutations approach presented two additional challenges. First, a number of instances were observed where the rate of mutations increased but did not show increase in reported infections (false positive signal). Second, the nature of incoming genomic data is generally noisy (due to smaller number of samples and weighting of different mutations shows large variations) and changes quickly, therefore, the ongoing most recent rate of mutations is very noisy as well. It was concluded that at this stage, rate (derivative) of mutations is not a reliable signal for surge prediction. In the future, this could be revisited with more stable reporting of genomic sequences with shorter sample collection to sequence publication timeframes. Figure 1-figure supplement 4 presents side by side comparison of the metrics investigated. Overall, it appears that collective non-synonymous mutations (k_a) provides the most reliable signal for surge prediction. In the remaining text, we discuss the key results and their importance.

Spike Protein: Spike protein interacts with the angiotensin-converting enzyme 2 receptor and plays a vital role in infecting the human cells.[11] Spike protein has been the target of mRNA-based vaccines. Viral sequences show significant changes in synonymous and non-synonymous mutations in the spike protein (188,166 unique sequences observed so far), with large increases ahead of the surge in reported human infections, most noticeably with the surges associated with the Gamma/Delta and the Omicron variants (Figure 1A). It is important to note that the mutations show increase 10-14 days before the increase in human infections. It is also interesting to note that the synonymous mutations (data available on the website) show decrease post surges. The decrease in mutations prior to the Omicron BA2 surge corresponds to reversal mutations (returning to reference sequence). However, at present the non-

synonymous and synonymous mutations post the Omicron variant remain elevated, more so than any period during the COVID19 outbreak.

Proteins showing significant mutations: In addition to the spike protein, SARS-CoV-2 membrane[12] (Figure 1B, 11,338 unique sequences observed so far) and envelope[13] (Figure 1C, 1,314 unique sequences observed so far) proteins have also shown significant mutations, starting just before the Omicron variant (November 2021 onwards). For the case of membrane protein, there was a significant increase that started in the Gamma/Delta variants (June 2021 onwards) and further increased just before the BA5 surge. The spike, membrane and envelope proteins are all located on the surface of SARS-CoV-2 and potentially interact with the components of the immune system. The large increase in mutations in all these external proteins assumes importance in post-vaccination period (discussed further below).

Other proteins: For comparison, Figure 1D shows mutations from RNA-dependent RNA polymerase (RDRP, 60,575 unique sequences observed so far), which has been targeted for development of antiviral drug therapies. Till present, RDRP has shown comparatively lower magnitude of non-synonymous mutations. Note that gray dots are individual mutations, the mean (black line) is weighted by number of sequences for each day by the mutations. Significant increases in mutations are also observed in NSPs 1, 4, 6, 13, 15, ORFs 6, 7a and 7b (data available on website). Overall, this analysis allows us to monitor ongoing mutations in different proteins; when rapid rise is observed over a short period of time, we issue surveillance watches and warnings (reserved for most extreme cases) for new possible variants with combination of proteins showing new mutations.

Vaccination and mutational frequencies: Wide-spread vaccination against SARS-CoV-2 (December 2020 onwards) coincides with significant increase in mutation rates of several SARS-CoV-2 proteins. Spike, membrane and envelope proteins have shown rapid mutations in especially in the Omicron variant (gray dots in Figure 1). This is possibly due to viral adaptations under the selective pressure exerted by the vaccine, as a significant number of mutations were observed in 2021, especially for the spike protein (gray dots in Figure 1A indicate spike protein has mutated much more than any other protein). The long-term effectiveness of mRNA-based SARS-CoV-2 vaccines remains unknown. After the initial regimen of two doses, the administration of additional booster (third and fourth) doses has decreased due to improvement in COVID19 fatality rates as well as political reasons.[14] This situation raises concerns. Other proteins have shown reversal mutations (higher similarity with the reference sequence) after periods of significant increase in mutations, however, post-vaccination the significant mutations observed in the spike, envelope and membrane protein related to the Omicron variant remain at extremely elevated levels. As Omicron, BA2, BA5 and subsequent variants are showing increased rates of transmission, gain or improvement of function in other proteins could lead to emergence of newer variants of concern. Over long-term this needs to be addressed by vaccines with longer periods of effectiveness and post-infection treatment options including antiviral drugs.

Surge prediction: The methodology presented here allows monitoring the potential increase in reported number of human infections. To date, spike protein has shown the most direct correlation in the rate of non-synonymous mutations and the rates of human infection. In particular, the case of Omicron variant and also the Gamma variant, spike protein showed rapid increase in mutations about 10-14 days ahead of time. Furthermore, membrane protein showed rapid mutations before surge related to BA5. Therefore, such increase in mutations serve as an indication of upcoming surges. For example, we issued

a surge watch on the website on June 29th 2022, which was converted to a warning on July 14th. This was confirmed by increase in infection cases worldwide throughout July (see Figure 1-figure supplement 5). Further, we issued an additional warning on September 7th 2022, which was confirmed by surge in several European countries including France, United Kingdom, Germany and Italy (see Figure 1-figure supplement 6).

The role of different (or dominant) SARS-CoV-2 variants in major surges is unclear at this time and needs further research. Different variants have been prevalent in different geographic regions at different times over the course of COVID19 outbreak, therefore, it is difficult to assign the surges to individual variants. In particular, Gamma and Delta variants were both prevalent in different countries in 2021. We are working on enabling this analysis by geographic locations and the results will be available through the website. However, at present our analysis is able to make predictions about collective surges before they occur, as illustrated by the case of BA5.

In the future, a number of factors could affect the performance of the presented approach. In particular, as the pandemic situation has improved in the second half of 2022, the number of tests being performed and the sequences being deposited into public repositories have decreased. Furthermore, it is widely being discussed that the population is showing increased immunity against the virus due to vaccination and naturally acquired immunity. The presented approach is dependent on availability of sequences, therefore, we hope that scientific community will continue to urge the medical community and public health agencies to commit resources to sequencing the positive COVID19 patients on a regular basis. Nonetheless, even with availability of smaller number of sequences, our approach is weighted by mutations and percentage of sequences showing non-synonymous mutations. Therefore, whenever new mutations show up in large percentages, our approach will still be able to work. On the other hand, viruses continue to evolve and if the population acquires large scale immunity leading to drastic reduction in number of infections, our surveillance approach would still allow preparation in cases of significant viral genome changes (such as going from SARS-CoV to SARS-CoV-2) whenever they occur and lead to the possibility of another major breakout.

DISCUSSION

The methodology and the website described here provides real time mutational changes of 26 SARS-CoV-2 proteins and ORFs. The changes in non-synonymous mutations correlate with the increase in reported cases of infections. Apart from identifying mutations of concern for in-depth scientific studies, the website is intended to keep the medical community informed about potential upcoming surges. Warnings of increase in mutations and expected surges are displayed on the website (and also available through email alerts). It should be noted that this real time analysis is dependent on the various health labs and medical facilities for swiftly depositing the viral genome sequences into the public databases such as the GenBank. The shorter the lag time in depositing the sequences by the wider community, more accurate and effective the prediction capabilities of our approach and the website will be.

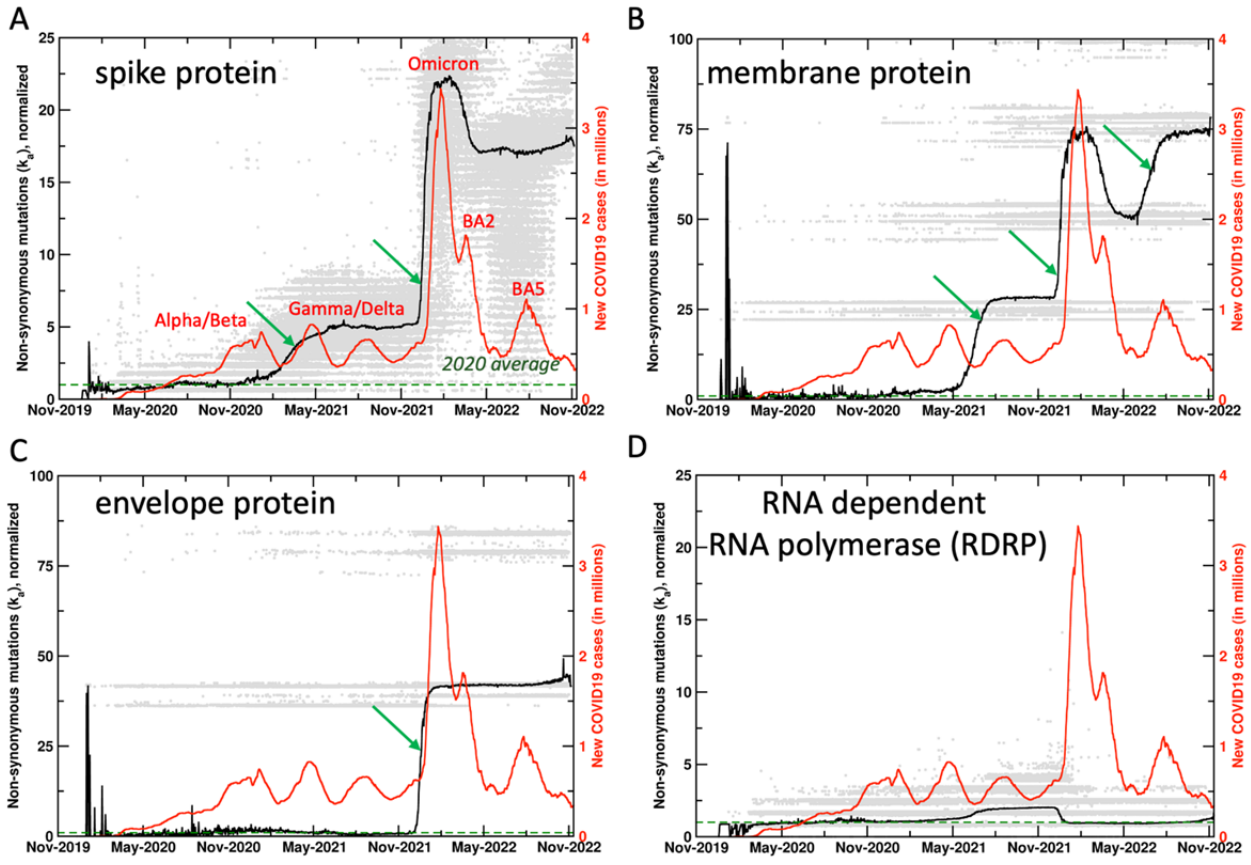


Figure 1: Mutations in SARS-CoV-2 proteins increase before COVID19 surges. Non-synonymous mutations over the course of the COVID19 outbreak were identified by analysis of 6.4 million sequences. Gray dots indicate individual mutations, while black lines show weighted means for each day. Red lines show new COVID19 cases (averaged weekly) across the world. The green arrows mark the time when new mutations occurred in significant numbers before the outbreaks, allowing prediction of future outbreaks. The mutation values have been normalized using average of all mutations in the year 2020 (the first full year of the pandemic) as 1 (marked by dashed lines). Raw results are available in figure supplement 1. Values of 0 indicate same sequence as the Wuhan sequence, while larger values indicate more mutations. Note that each gray dot corresponds to a unique sequence, and there can be multiple sequences showing the same mutations. The weighted mean for the day is calculated by using all sequences reported for the day. The peaks for COVID19 cases are labeled with prevalent variants. Alpha/Beta, Omicron and Omicron BA2, BA5 were the prevalent variants at the time of 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.

Legends for the supplement figures

Figure 1-figure supplement 1: Un-normalized results for the mutations in SARS-CoV-2 proteins. See Figure 1 in the main manuscript for more details. Here the raw results for the four proteins are plotted for the non-synonymous mutations. The same y-axis scale is used for comparison of the mutations across all the four proteins shown.

Figure 1-figure supplement 2: Ratio of non-synonymous mutations/synonymous mutations in SARS-CoV-2 proteins. The commonly used indicator did not provide a reliable signal for surge prediction for most proteins. The ratio for only the membrane protein shows increase before the surges associated with some variants. The information from this ratio can be used as a secondary signal to support the primary signal from k_a.

Figure 1-figure supplement 3: Daily rate of non-synonymous mutations in SARS-CoV-2 proteins. The rate is calculated as a numerical derivative of data shown in Figure S1. The rate shows most noticeable increase before the Omicron surge, other periods are inconclusive. Note that the nature of ongoing current data is expected to be noisy (few samples, weightings that change over days), therefore, the rate of mutations appears to be unreliable in predicting surges.

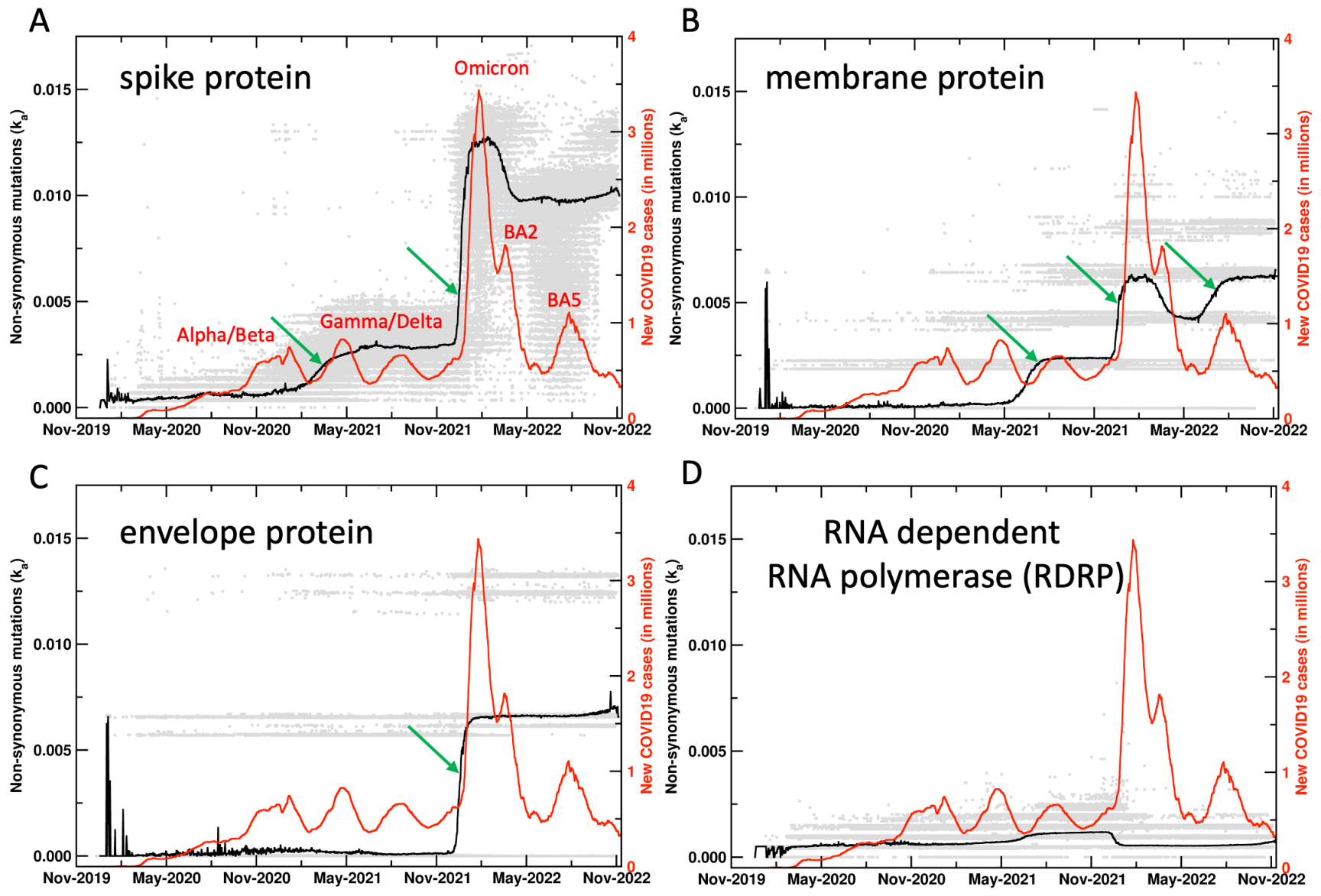
Figure 1-figure supplement 4: Side by side comparison of various metrics considered in this study. The quantities shown here are same as depicted in Figure 1 (main manuscript) and Figure S2 and S3. See the legends of other figures for details.

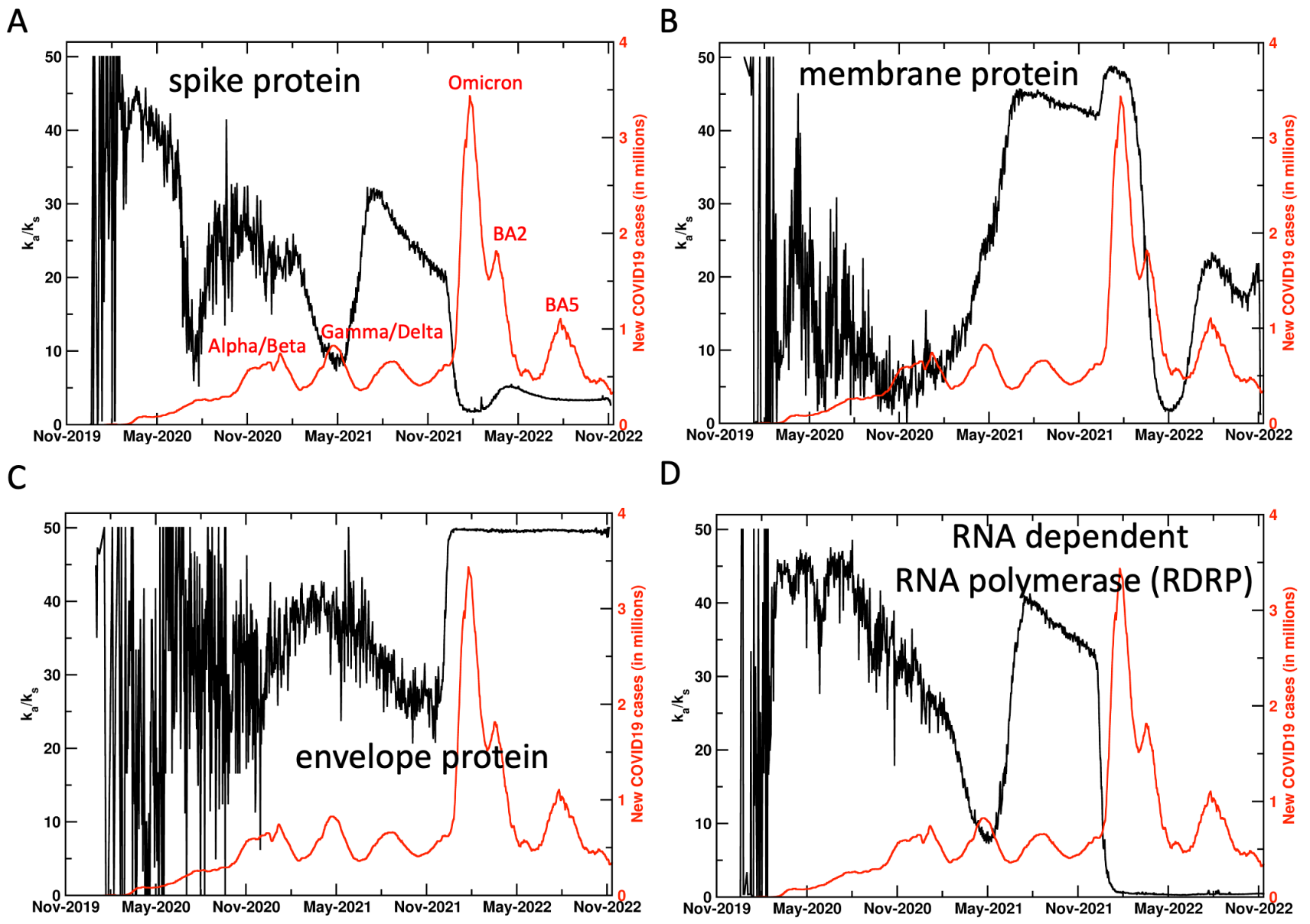
Figure 1-figure supplement 5: Performance of the surge watch and warning issued on June 29th 2022 and July 14th 2022 respectively. The number of infection cases showed a sustained increase after the issue of our watch on June 29th. This watch was elevated to a warning on July 14th (Note, warning is considered more severe than a watch) and the number of cases showed a further increase. The warning was removed on August 30th 2022. The number of cases peaked roughly a month after our watch was issued.

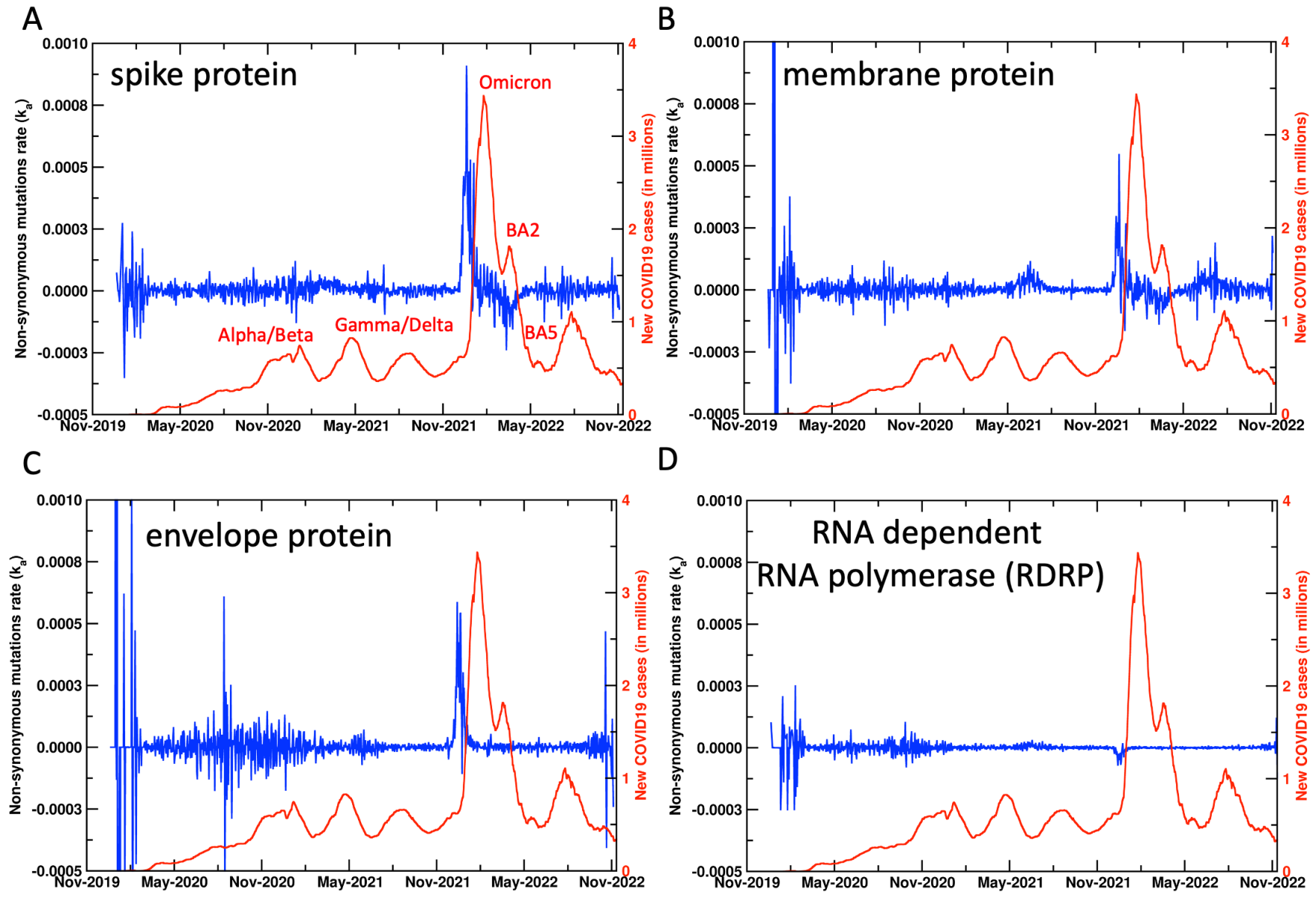
Figure 1-figure supplement 6: Performance of the surge watch issued on September 7th 2022. The number of infection cases showed a sustained increase in Europe and several individual countries, after we issued our watch. The data did not warrant a further elevation and this watch was eventually removed on November 14th 2022.

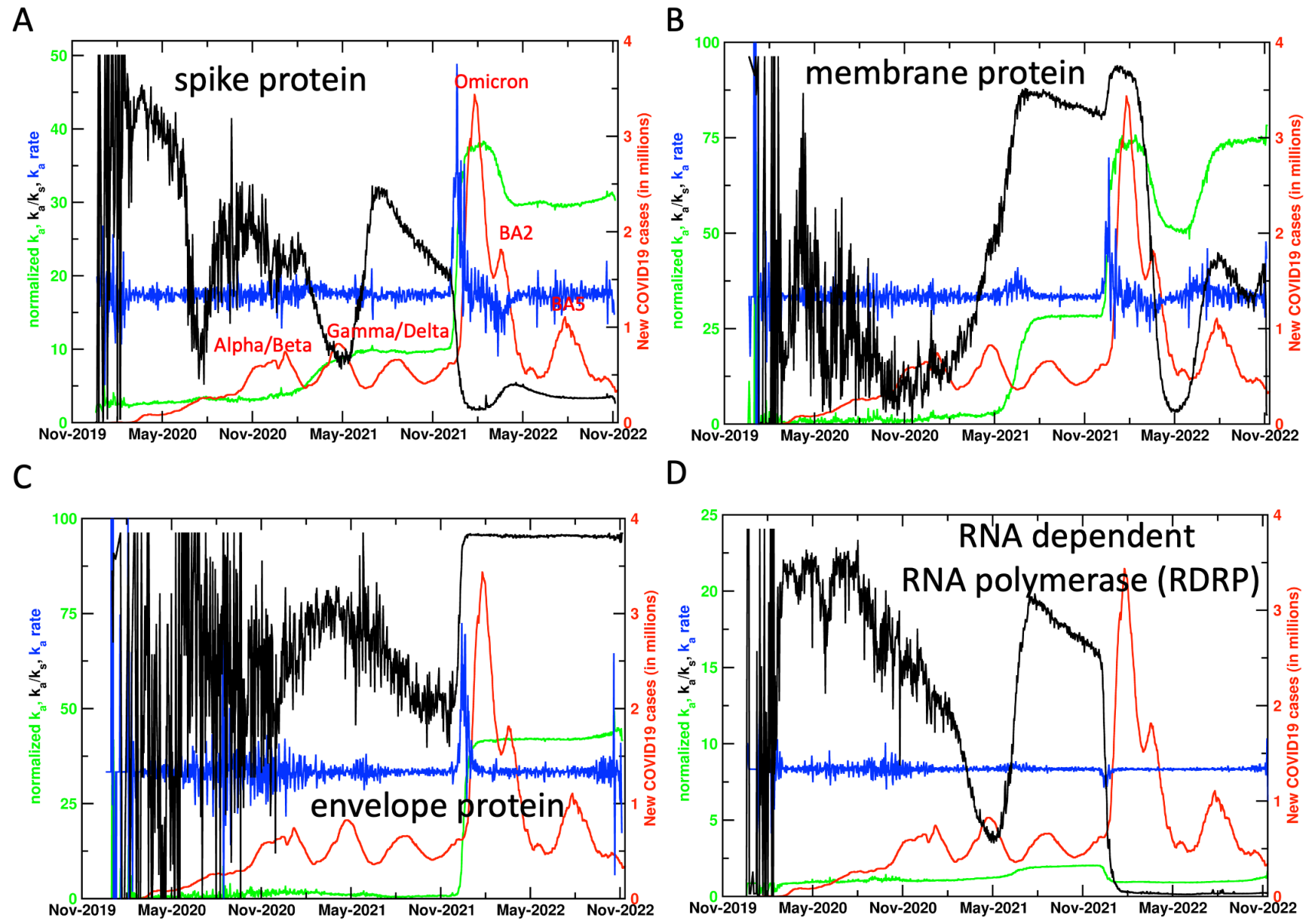
References

1. Platto, S, Wang, Y, Zhou, J, Carafoli, E. History of the COVID-19 pandemic: Origin, explosion, worldwide spreading. Biochemical and Biophysical Research Communications, 2021;538;14-23.
2. Zhang, Z, Li, J, Zhao, XQ, Wang, J, Wong, GKS, Yu, J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics, Proteomics and Bioinformatics 2006;4(4);259-263.
3. Kistler KE, Huddleston J, Bedford T. Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. Cell Host & Microbe. 2022;30;545-555.
4. Dong, E, Du, H, Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet infectious diseases 2020;20(5);533-534.
5. Nadkarni, PM, Parikh, CR. An eUtils toolset and its use for creating a pipeline to link genomics and proteomics analyses to domain-specific biomedical literature. J Clin Bioinforma. 2012;2(1);9.
6. Sievers F, Higgins DG. Clustal omega. Curr Protoc Bioinformatics 2014;48;3.13.1-16.
7. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 2006;34;W609-12.
8. Zhang Z. KaKs_calculator 3.0: Calculating selective pressure on coding and non-coding sequences. Genomics Proteomics Bioinformatics 2022;S1672-0229(21)00259.
9. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994;11(5);725-36.
10. Wu, F, Zhao, S, Yu, B, Chen, YM, Wang, W, Song, ZG, Hu, Y, Tao, ZW, Tian, JH, Pei, YY, Yuan, ML. A new coronavirus associated with human respiratory disease in China. Nature 2020; 579(7798);265-269.
11. Xia, X. Domains and functions of spike protein in Sars-Cov-2 in the context of vaccine design. Viruses 2021;13(1);109.
12. Lu, S, Ye, Q, Singh, D, Cao, Y, Diedrich, JK, Yates, JR, Villa, E, Cleveland, DW, Corbett, KD. The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. Nature Communications 2021;12(1);1-15.
13. Zheng, M, Karki, R, Williams, EP, Yang, D, Fitzpatrick, E, Vogel, P, Jonsson, CB, Kanneganti, TD. TLR2 senses the SARS-CoV-2 envelope protein to produce inflammatory cytokines. Nature Immunology 2021;22(7);829-838.
14. Sabahelzain, MM, Hartigan-Go, K, Larson, HJ. The politics of COVID-19 vaccine confidence. Current Opinion in Immunology. 2021;71;92-6.


A
spike protein
Omicron
BA2
BA5
Alpha/Beta
Gamma/Delta
B
membrane protein
C
envelope protein
D
RNA dependent
RNA polymerase (RDRP)
Non-synonymous mutations (k_a)
New COVID19 cases (in millions)
0.015
0.010
0.005
0.000
4
3
2
1
0
Nov-2019
May-2020
Nov-2020
May-2021
Nov-2021
May-2022
Nov-2022


A
spike protein
Omicron
BA2
BA5
Alpha/Beta
Gamma/Delta
B
membrane protein
C
envelope protein
D
RNA dependent
RNA polymerase (RDRP)
k_a/k_s
New COVID19 cases (in millions)
Nov-2019
May-2020
Nov-2020
May-2021
Nov-2021
May-2022
Nov-2022


A
spike protein
Omicron
BA2
BA5
Alpha/Beta
Gamma/Delta
B
membrane protein
C
envelope protein
D
RNA dependent
RNA polymerase (RDRP)
Non-synonymous mutations rate (k_a)
New COVID19 cases (in millions)
0.0010
0.0008
0.0005
0.0003
0.0000
-0.0003
-0.0005
4
3
2
1
0
Nov-2019
May-2020
Nov-2020
May-2021
Nov-2021
May-2022
Nov-2022

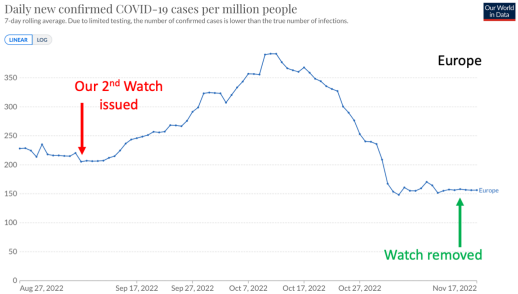

A
spike protein
Omicron
BA2
BA5
Alpha/Beta
Gamma/Delta
B
membrane protein
C
envelope protein
D
RNA dependent
RNA polymerase (RDRP)
normalized k_a, k_a/k_s, k_a rate
New COVID19 cases (in millions)
Nov-2019
May-2020
Nov-2020
May-2021
Nov-2021
May-2022
Nov-2022

Daily new confirmed COVID-19 cases per million people

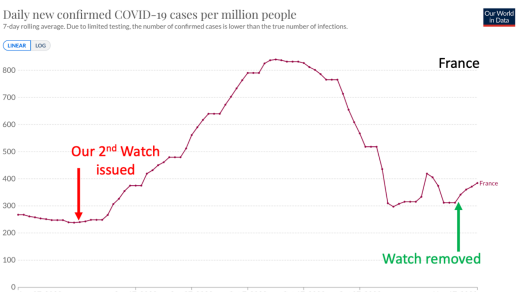
7-day rolling average. Due to limited testing, the number of confirmed cases is lower than the true number of infections.

Our World in Data

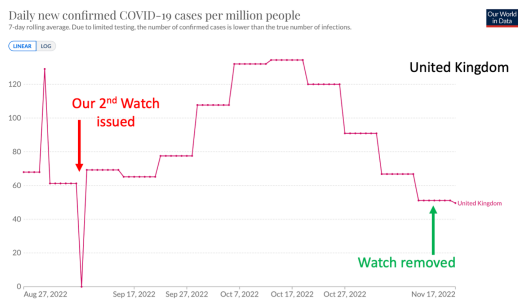
Source https://ourworldindata.org/coronavirus



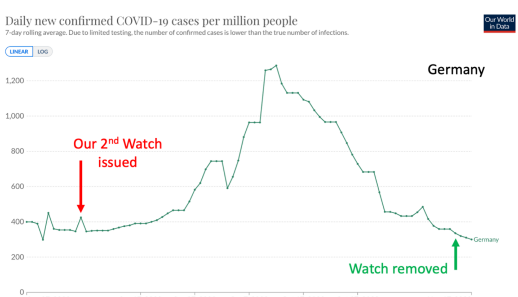
Source https://ourworldindata.org/coronavirus



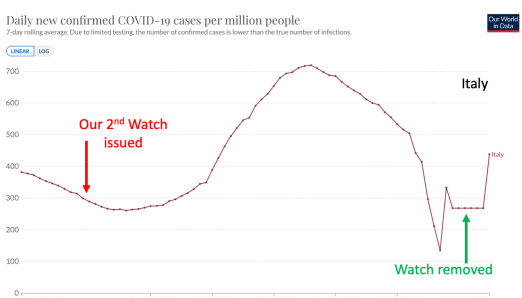
Source https://ourworldindata.org/coronavirus



Source https://ourworldindata.org/coronavirus



Source https://ourworldindata.org/coronavirus



Source https://ourworldindata.org/coronavirus