Bioinformatics, 38(14), 2022, 3549–3556
https://doi.org/10.1093/bioinformatics/btac370
Advance Access Publication Date: 31 May 2022
Original Paper

OXFORD

Sequence analysis

VOC-alarm: mutation-based prediction of SARS-CoV-2 variants of concern

Hongyu Zhao[1,†], Kun Han[2,†,‡], Chao Gao[3,4,†], Vithal Madhira[5], Umit Topaloglu[1,6,7], Yong Lu[2,6,*,‡] and Guangxu Jin [1,6,*]

[1]Department of Cancer Biology, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA, [2]Department of Microbiology and Immunology, Wake Forest School of Medicine, Winston-Salem, NC 27101, USA, [3]Department of Gynecology and Obstetrics, Tianjin Medical University General Hospital, Tianjin, China, [4]Tianjin Key Laboratory of Female Reproductive Health and Eugenics, Tianjin 300052, China, [5]Palila Software LLC, Reno, NV 89521, USA, [6]Wake Forest Baptist Comprehensive Cancer Center, Winston-Salem, NC 27157, USA and [7]Wake Forest School of Medicine, Center for Biomedical Informatics, NC 27101, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.
‡Present address: Department of Cancer, Houston Methodist Hospital, Houston, TX 77030, USA
Associate Editor: Alfonso Valencia

Received on January 20, 2022; revised on April 3, 2022; editorial decision on May 23, 2022; accepted on May 26, 2022

Abstract

Summary: Mutation is the key for a variant of concern (VOC) to overcome selective pressures, but this process is still unclear. Understanding the association of the mutational process with VOCs is an unmet need. Motivation: Here, we developed VOC-alarm, a method to predict VOCs and their caused COVID surges, using mutations of about 5.7 million SARS-CoV-2 complete sequences. We found that VOCs rely on lineage-level entropy value of mutation numbers to compete with other variants, suggestive of the importance of population-level mutations in the virus evolution. Thus, we hypothesized that VOCs are a result of a mutational process across the globe. Results: Analyzing the mutations from January 2020 to December 2021, we simulated the mutational process by estimating the pace of evolution, and thus divided the time period, January 2020—March 2022, into eight stages. We predicted Alpha, Delta, Delta Plus (AY.4.2) and Omicron (B.1.1.529) by their mutational entropy values in the Stages I, III, V and VII with accelerated paces, respectively. In late November 2021, VOC-alarm alerted that Omicron strongly competed with Delta and Delta plus to become a highly transmissible variant. Using simulated data, VOC-alarm also predicted that Omicron could lead to another COVID surge from January 2022 to March 2022.
Availability and implementation: Our software implementation is available at https://github.com/guangxujin/VOC-alarm.
Contact: gjin@wakehealth.edu or ylu2@houstonmethodist.org
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

To track the mutations in the SARS-CoV-2 variants, genomic surveillance has been made by the Global Initiative on Sharing All Influenza Data (GISAID) (Elbe and Buckland-Merrett, 2017), Nextstrain (McBroome *et al.*, 2021) and Phylogenetic Assignment of Named Global Outbreak (Pango) (Graham, 2020). To better describe the variants of potential global health significance, the World Health Organization (WHO) has established an international surveillance system to designate variants into different categories, like variant of concern (VOC) (Sheikh *et al.*, 2021; Sonabend *et al.*, 2021; Wang *et al.*, 2021) or variant of interest (VOI) (Chakraborty *et al.*, 2021; Thompson *et al.*, 2021; Wang *et al.*, 2021). The VOCs, including Alpha (Zhang *et al.*, 2020), Beta (Hung *et al.*, 2020), Gamma (Collier *et al.*, 2021) and Delta (Del Rio *et al.*, 2021), have shown increased transmissibility and disease severity. On November 26, 2021, the WHO designated the variant B.1.1.529 as another VOC, named Omicron (Chen *et al.*, 2021). The transmissibility and caused disease severity of Omicron remained unclear.

Genomic surveillance provided valuable genomic mutation information to further understand these VOCs (Saito *et al.*, 2021). Great efforts have been made to construct models, which aimed to

understand the transmissibility of SARS-CoV-2 (Davies *et al.*, 2021; Dhar *et al.*, 2021; Jentsch *et al.*, 2021) and the SARS-CoV-2 pandemic (Davies *et al.*, 2021; Dhar *et al.*, 2021; Kontis *et al.*, 2020; Narykov *et al.*, 2021; Sonabend *et al.*, 2021). However, due to the complexity of the mutation data, including lineages, clades and related geographical locations, understanding how the mutations determined the VOCs and how the VOCs could compete with other variants has been difficult. To fully make advantage of the valuable mutation information, we must find out whether the mutational process of a VOC has been driven by a force from selective pressures, which were caused by diagnostic approaches, treatments and/or vaccines (Choi *et al.*, 2021; Collier *et al.*, 2021; Lopez Bernal *et al.*, 2021; Payne *et al.*, 2021; Sheikh *et al.*, 2021; Wilder-Smith and Mulholland, 2021).

It remains unclear that how the mutations determined VOCs. Here, we developed a novel computational method, VOC-alarm, to address the association of the increased mutations with VOCs and their caused COVID surges. Our findings confirmed that Omicron variant would be a new VOC, however, it is still in its early stage in which it has been competing with Delta and Delta plus variants, which might continue into January 2022. The Omicron's caused surge would follow the current one related to Delta and Delta plus and will reach its peak no later than March 2022.

2 Results

2.1 Competition among variants sponsored by speed-up mutation led to VOCs

To simulate the force that has driven VOCs and understand how mutation could determine VOCs, we identified the changes in the pace of evolution by mutations. We first analyzed the mutation information from GISAID (Elbe and Buckland-Merrett, 2017), including 91 382 343 whole-genome mutations from 3 816 807 complete SARS-CoV-2 sequences, as of September 28, 2021 (Supplementary Table S1). Despite classified lineages or clades defined by GISAID (Elbe and Buckland-Merrett, 2017), Nextstrain (McBroome *et al.*, 2021) and Pango (Graham, 2020), as well as the diverse geographical locations included in the mutation information, current VOCs have not been dominated by any known lineages and their emergence processes have not been restricted within certain geographic locations. To fully make use of the complex mutation information and understand its association with VOCs, we must reconsider the mutational process of the virus as a global behavior that was not limited by known lineages and clades so that we could predict any emerging lineage in an unexpected geographical region. To accomplish this goal, we estimated the mutation rate across the globe by a spatiotemporal genomic variation (SGV) index (Section 4, Supplementary Fig. S1 and Supplementary Table S2) and its weekly change rate as the pace of evolution (Fig. 1A).

SGV, calculated as an average mutation number per genome sequence across the globe, showed a continuous increase. It increased to 30 by end September 2021 (Supplementary Table S3 and Supplementary Fig. S2), suggesting that the mutation rate was ~5 mutations/100 days. Moreover, the increased rate periodically changed between 6% and 11% (Fig. 1A, top, Supplementary Table S4), which separated the time period from January 1, 2020 to September 28, 2021 into Stages I–V. Remarkably, Stages I and III with accelerated mutation speeds included a small number of Alpha/B.1.1.7 variants (Chemaitelly *et al.*, 2021; Payne *et al.*, 2021; Washington *et al.*, 2021) (Fig. 1B and C, Supplementary Fig. S3 and Supplementary Tables S5 and S6) and Delta variants [including AY.4 and B.1.617.2 lineages (Celik and Tallei, 2022; Lopez Bernal *et al.*, 2021; Pung *et al.*, 2021; Sonabend *et al.*, 2021)] (Fig. 1B, right, Supplementary Fig. S4 and Supplementary Table S7). Despite the tiny populations, strong competitions have been observed for the collected sequences of Alpha and Delta variants (Fig. 1A, lower). To clarify the competition sponsored by speed-up mutations, we simulated the adaptiveness to selective pressures (Choi *et al.*, 2021; Collier *et al.*, 2021; Lopez Bernal *et al.*, 2021; Payne *et al.*, 2021; Sheikh *et al.*, 2021; Wilder-Smith and Mulholland, 2021) by the mutational

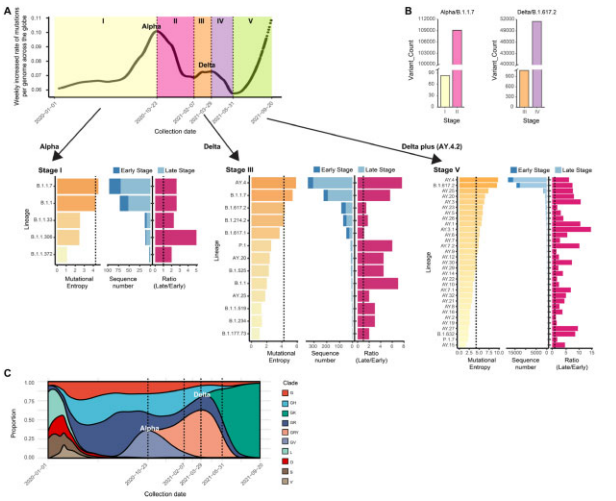

Fig. 1. Determining VOCs from their competitions with other variants. (**A**) Competition among variants sponsored by speed-up mutation led to VOCs. Top: mutation changes classified the time period from January 2020 to September 2021 into Stages I–V, dash lines: peaks or valleys. Shown is the weekly mutation number change rate per genome across the globe, which is used to evaluate the pace of evolution (Section 4). Below: the competitions among variants in Stages I, III and V sponsored by speedy-mutation changes. The competition was evaluated by the mutational entropy value of the mutation numbers of the genomes form each lineage (left), which simulated the adaptiveness of a variant to the selective pressures. The growth rate of each variant in each stage was shown by the sequence numbers of the late and early periods of each stage, i.e. early and late stages (middle) and their ratio (right). (**B**) Numbers of Alpha sequences in Stages I and II (left) and those of Delta sequences in Stages III and IV (right). (**C**) Stream plot of the proportions of the GISAID clades across Stages I–V. Shown is the percentage of the sequences collected on a specific date for a specific clade

entropy concept (Fan *et al.*, 2021, 2022; Fariselli *et al.*, 2021; Mukherjee *et al.*, 2013) (Section 4). Of five candidate lineages with population growths and significantly high mutation numbers (Fig. 1A, lower left, Supplementary Fig. S5 and Supplementary Table S8), lineage B.1.1.7 showed the highest mutational entropy (Fig. 1A, lower left and Supplementary Figs S5 and S6). To illustrate the difference among mutational entropy, mutation number and population growth rate, we tested whether mutation number or population growth rate could identify the VOC in Stage I. However, lineage B.1.1.7 only ranked at top by mutational entropy but not mutation number and population growth rate (Supplementary Fig. S7). This suggested that mutational entropy described the competition among the variants best based on the adaptiveness to selective pressures. Similarly, of the 13 lineages with population growths and significantly high mutation numbers, identified from Stage III, we ranked the lineages AY.4 and B.1.617.2 at top by mutational entropy values (Fig. 1A, lower middle, Supplementary Fig. S8 and Supplementary Table S9). Thus, we identified Alpha and Delta variants by their competitions with other variants in Stages I and III. To better describe the competitive capabilities by mutational entropy, we used the mutational entropy value of lineage B.1.1 in Stage I as the threshold for identifying lineages as VOCs.

Unexpectedly, Stage V also showed an accelerated pace of evolution, which predicted the essential role of AY.4 variant in October 2021 (Fig. 1A, lower right, Supplementary Fig. S9 and Supplementary Table S10). In contrast, Stages II and IV with the decelerated paces were related to the COVID surges caused by Alpha and Delta variants. These two stages included the dates of the official designation for Alpha and Delta variants by the WHO (Supplementary Table S11). Strikingly, Stage V with the accelerated mutation speed generated a new generation of Delta variant, Delta plus (AY.4.2 sublineage) (Angeletti *et al.*, 2021), evolved from AY.4 lineage and emerged in October 2021.

2.2 Competition-driven prevalence change

Consistent with the ranking of lineages by mutational entropy, AY.4 variant (Delta) in Stage V with the highest mutational entropy

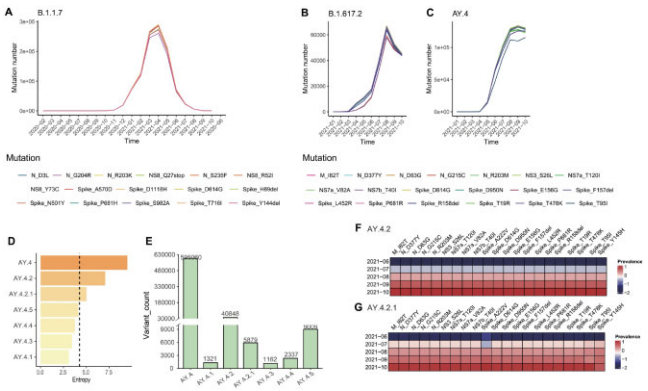

Fig. 2. Competition-driven sequence prevalence change. The prevalence lineplots of the mutations of the VOCs by months. B.1.1.7 (**A**), B.1.617.2 (**B**) and AY.4 (**C**). Color: scaled sequence prevalence. Columns: Marker mutations for the VOC definitions. The top 50 amino acid mutations sorted by sequence prevalence were shown in Supplementary Figure S12. (**D**). Mutational entropy values for AY.4 sublineages in October 2021. (**E**) Sequence counts for the AY.4 sublineages as of November 18, 2021. (**F and G**) The prevalence heatmaps of the mutations of AY.4.2 (F) and AY.4.2.1 (G) by months. Color: scaled sequence prevalence

showed a significant prevalence increase in October 2021, whereas B.1.1.7 (Alpha) and B.1.617.2 (Delta) variants showed a decrease trend (Fig. 2A–C). The VOC (AY.4) showed a 12% accumulative prevalence worldwide and about 40% prevalence in the sequences collected from the UK in October and early November (Supplementary Fig. S10). Distinctly, B.1.1.7 and B.1.167.2 did not continue their increased prevalence trends in the worldwide GISAID sequences (Fig. 2A and B and Supplementary Figs S11 and S12). These results indicated that the competition sponsored by mutation could lead to the population change of the VOCs.

We observed not only inter-lineage but also intra-lineage competitions. Analyzing the mutational entropy values of AY.4 sublineages in October 2021, we found that both AY.4.2, which was also called Delta plus, and AY.4.2.1 sublineages emerged as VOCs (Fig. 2D). AY.4.2 variant was suspected to be a potential VOC (Arora *et al.*, 2022; Saunders *et al.*, 2022), which has two characteristic amino acid mutations in the Spike protein, Y145H (Aljindan *et al.*, 2021) and A222V (Kannan *et al.*, 2021), closely monitored by the WHO and the CDC in late October. This finding confirmed our prediction in Stage V (as of September 2021), regarding the important role of lineage AY.4 in the future pandemic. Interestingly, AY.4.2 and AY.4.2.1 variants displayed a strong increase in prevalence in October (Fig. 2E–G), compared to AY.4 variant.

Further analysis of the sequences of AY.4.2 across the globe found a significant mutation number increase in October 2021 (Fig. 3A), reaching to 100. AY.4.2 has been found in 45 countries from different continents (Fig. 3B–D), which led to an increase in cases or deaths in October and November 2021 in these countries (Fig. 3E–G). These results indicated that competition among variants could lead to the variant population change and COVID surge.

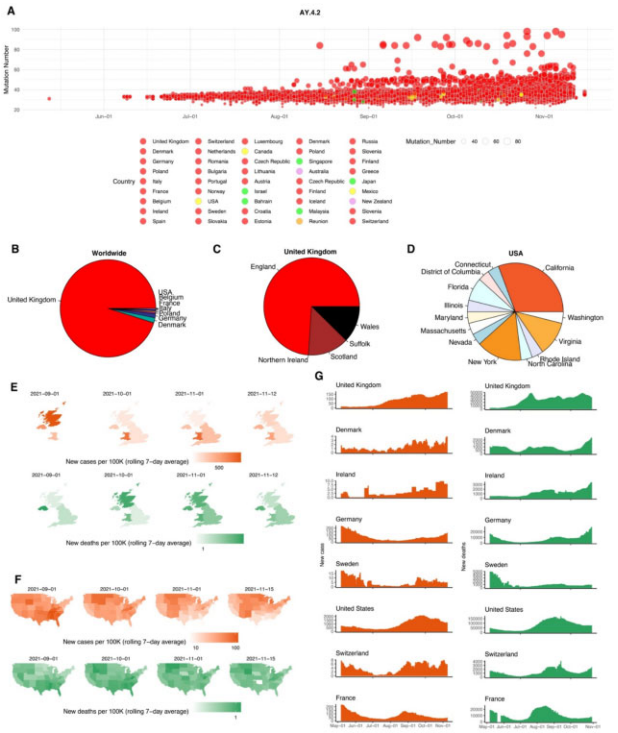

Fig. 3. Mutation number, geographical distribution, and associated COVID surges of AY.4.2 variant. (**A**) Scatter plot of AY.4.2 variants by their mutation numbers and collected dates. Color: the continent in which the sequence was collected. (**B–D**) Distribution of AY.4.2 variants in the worldwide (B), UK (C) and USA (D). Color: the country in which the sequence was collected. (**E** and **F**) Cases and deaths surges in different regions of the UK (E) and USA (F). Color: rolling 7-day average numbers of new cases (top) and new deaths (lower). (**G**) COVID surges in the countries in which AY.4.2 sequence was found in October or November 2021

2.3 Omicron emerged as the only variant that could compete with Delta in November

Unexpectedly, following the emergence of Delta plus, VOC-alarm identified another stage (VII) with an accelerated pace of evolution and predicted the B.1.1.529 variant as the first lineage that showed a relatively high mutational entropy except of the known VOCs (Fig. 4A–C, Supplementary Fig. S13 and Supplementary Table S12). This result suggested that B.1.1.529 has the potential to be next VOC. As of December 7, 2021, 12 of 21 Delta variants took part in the competitive process in Stage VII, in which AY.4, AY.4.2, AY.34, AY.120, AY.4.2.1 and B.1.617.2 increased their total prevalence from <25% in August 2020 to >50% (Fig. 4D). Remarkably, despite a small population, B.1.1.529 variant had increased its prevalence from 0.0035% on November 8, 2021 to 0.6% on November 26, 2021 (Fig. 4E and Supplementary Fig. S14). Of note, the speedy-mutating and fast-growing variant, B.1.1.529, was still in its early stage, requiring more time to develop into a VOC.

2.4 Future trend of the virus evolution

To figure out how long Omicron might take to complete its competition with other variants, we used the estimated pace of evolution data to simulate the future evolution of the virus. The regression on the pace of the evolution suggested that the selective process determined by the competition between Omicron and Delta/Delta plus would continue to January 2022, i.e. the end of simulated Stage VII, and its caused COVID surge would happen between January and March 2022, i.e. within the simulated Stage VIII (Fig. 5A). The total days for the competition between Omicron and Delta/Delta plus might take about 90 days

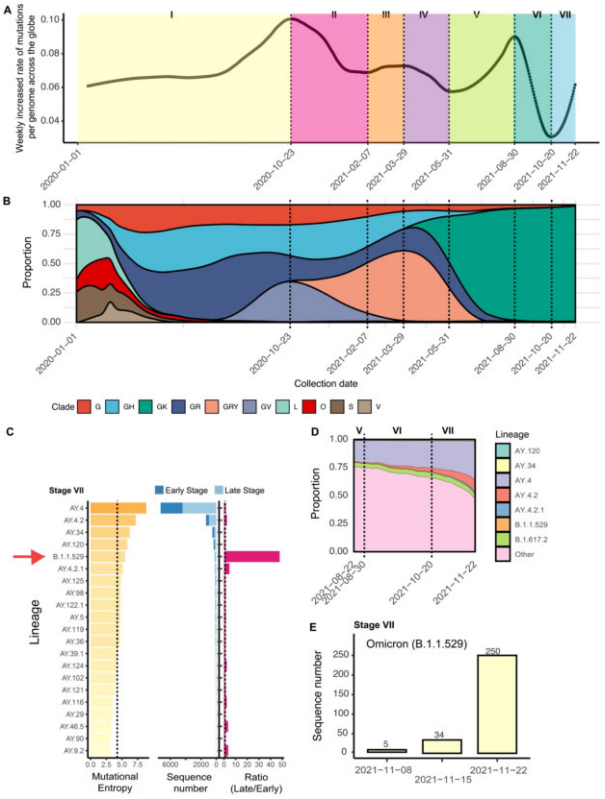

Fig. 4. Omicron emerged as the only variant competing with Delta variant. (**A**) Stages I–VII identified by the pace of evolution using the GISAID data as of December 7, 2021. Shown is the weekly mutation number change rate per genome across the globe. (**B**) Stream plot of the proportions of the GISAID clades across Stages I–VII. Shown is the percentage of the sequences collected form a specific clade on a specific date. (**C**) Competition among the variants in Stage VII evaluated by mutational entropy. The growth rate of each variant in each stage was shown by the sequence numbers of the late and early periods of each stage, i.e. early and late stages (middle) and their ratio (right). (**D**) Stream plot of the proportions of the top lineages in (C). (**E**) Increase of the collected Omicron sequence number in Stage VII

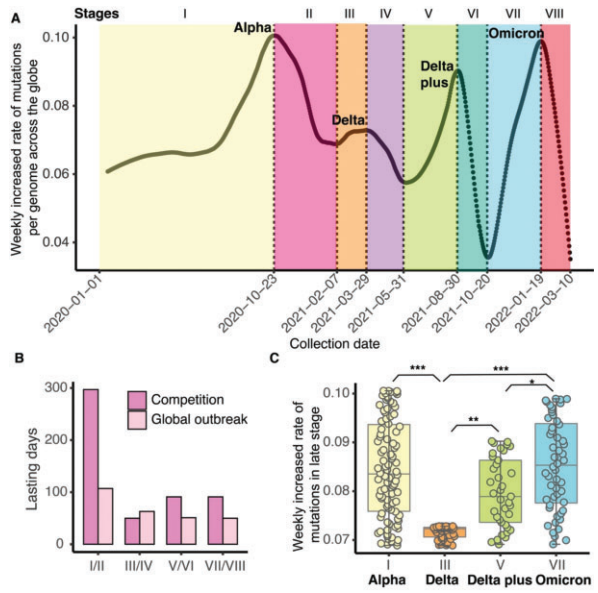

Fig. 5. Omicron may cause a COVID surge from January to March 2022. (**A**) Stages I–VIII. The late period of Stages VII and VIII were simulated by existing data of the pace of evolution and the trend from Delta plus. (**B**) The lasting days for different stages. (**C**) The comparisons of the paces of evolution among Stages I, III, V and VII. *<0.01, **<0.001 and ***<0.0001

and that for its global outbreak might take about 50 days (Fig. 5B).

We compared the paces of the evolution that caused the speedy mutation of the VOCs in Stages I, III, V and VII (predicted for Omicron). From Alpha to Delta, the pace of evolution was significantly decreased (Fig. 5C, $P < 0.0001$, Student's t-test), which might be related to the fast rollouts of vaccines in late 2020 and early 2021. However, from Delta to Delta plus and Omicron, the pace of evolution has been significantly increased (Fig. 5C, $P < 0.001$, Student's t-test). This might be associated with the adaptiveness of the new VOCs to the selective pressures caused by vaccines. As an instance, the first case in California, USA, infected by Omicron was fully vaccinated (Chen *et al.*, 2021), which suggested that Omicron might gain the adaptiveness to existing vaccines. Our results supported that Omicron might be a highly transmissible variant with high adaptiveness to vaccines, which may cause a new COVID surge earlier than March 10, 2022.

3 Discussion

Understanding the association between mutation and VOCs is urgently needed. A major breakthrough in this work is to simulate the competition among variants by mutation to answer questions related to VOCs: which lineage and when it will become a VOC and what likes its caused COVID surge.

We noticed that VOCs tend to grow from a small or even tiny population. In particular, Alpha, Delta and Omicron, remained a small number of sequences in Stages I, III and VII during they emerged. Remaining a small population should be more convenient for a VOC to become adaptive to the new selective pressures. As an evidence, we found that along with the emergence of VOCs, the precedent lineages decreased their population sizes. The B.1.1.7 precedent variants in GR clade and those of B.1.617.2 (Lopez Bernal *et al.*, 2021; Pung *et al.*, 2021; Sonabend *et al.*, 2021) in G clade significantly decreased their populations after lineages B.1.1.7 and B.1.617.2 emerged (Supplementary Fig. S15). These results suggested that speedy-mutation sponsored competition among variants could lead to prevalence change no matter how small the population size of the emerged VOCs was. In this work, we used the clades that showed most significant decrease in prevalence to identify VOCs. Similarly, we could identify Beta variant (B.1.351, Supplementary Fig. S16) from the clades with less statistical significance, which generally showed lower mutational entropy values. Since both Beta and Gamma (P.1, Fig. 1A) illustrated lower mutational entropy values, we did not designate them as VOCs in our alarming system.

Mutational entropy concept plus our defined stages displayed its power in predicting VOCs. In this article, we used 4.25, i.e. the mutational entropy value of lineage B.1.1 in Stage I, as the threshold to predict VOCs, which could be also applied to predict future VOCs. The mutational entropy, distinct from mutation number and population growth rate, was useful to identify the VOCs. Our defined mutational entropy was a metric to fully make advantage of the mutation information of the sequences from distinct lineages, different clades and various geographical regions. Development of mutational entropy concept for predicting VOCs was based on our novel modeling of the virus mutation as a global behavior. These modeling advances distinguished VOC-alarm from existing methods using genomic surveillance by phylogenetic clustering or those based on marker mutations (Ascoli, 2021; Harvey *et al.*, 2021; Muecksch *et al.*, 2021; Wang *et al.*, 2021). In fact, current VOCs were not limited by any known marker mutations, such as those from Spike protein (Arbeitman *et al.*, 2021), because of the unexpected speedy mutational process and fast gained mutations on all proteins and whole-genome regions. More importantly, it seems that the virus has been taking its mutational process as a strategy to overcome selective pressures; consequently, the mutational process was not limited to any marker mutations, any lineages, and any geographical locations.

Bearing that in mind, we could see that Stage VII did not reach its end yet in December 2021, suggesting that the virus continues an accelerated pace. The current small population of Omicron would

continue its fast evolution and speedy mutation to grow as a VOC. Since Omicron was at its early stage, another B.1.1 sublineage or an AY.4 sublineage was possible to compete with B.1.1.529. Similarly, we observed the same trend as B.1.1.529 from AY.4.2.1 sublineage (Supplementary Fig. S17). VOC-alarm would monitor the change in the pace of evolution and the mutational entropy of B.1.1.529 in real time.

Our VOC-alarm could monitor the pandemic and answer when it would terminate. In future, if the pandemic terminates, the pace of evolution would decrease to a low level, eventually terminating within a final stage; simultaneously, the mutational entropy values of all lineages would decrease to a lower level, e.g. much lower than 4.25. On the contrary, if the pandemic becomes an endemic, the virus might continue its pace in evolution, leading to periodically regional outbreaks (Benvenuto *et al.*, 2020).

VOC-alarm had three major conceptual advances in studying SARS-CoV-2 mutations: (i) reconsidering the virus mutational process as a global behavior. Distinct from phylogenetic tree analysis focusing on difference among variants, we fully made advantage of the sequence mutation information across the globe to analyze the time-lapse changes in the pace of evolution; (ii) estimating the competition among variants by our defined mutational entropy. Mutational entropy was a product of our novel thinking of the virus mutational process as a global behavior, which used the mutation information of a lineage across the globe; and (iii) defining stages by the pace of evolution. Stages separated the VOC emergence from its caused global outbreak, and thus enabled investigation of the competition of variants in a specific VOC emergence period. In summary, VOC-alarm would become an essential tool to predict VOCs and guide public health responses.

4 Materials and methods

4.1 SARS-CoV-2 mutation data

We used the SARS-CoV-2 mutations from global genomic surveillance. A total of 91 382 343 whole-genome mutations of 3 816 807 complete SARS-CoV-2 genomes from the GISAID database (https://www.gisaid.org/) as of September 28, 2021 (Elbe and Buckland-Merrett, 2017) were used to study Alpha and Delta variants (Supplementary Table S1). Another dataset of 156 205 744 whole-genome mutations from 5 709 730 complete SARS-CoV-2 genomes from the GISAID database (https://www.gisaid.org/) as of December 7, 2021 (Elbe and Buckland-Merrett, 2017) were used to analyze Delta Plus and Omicron variants (Supplementary Table S13). The meta data included information geographic location, clade, lineage, collection date and age and gender of confirmed cases. The latest data from outbreak.info (https://outbreak.info/) were used to analyze the prevalence of VOCs.

4.2 SGV index

The challenge in understanding how the virus mutates into a VOC is from that we know little about the process of the virus mutation. Because of separated sequences by lineages or clades and divided geographical regions, there lacks a global view of the evolution process of the virus. To enable an understanding of the evolution process across the globe and take the virus mutational process as a global behavior, we developed the SGV index, defined by the mean value of the mutation numbers of variants collected worldwide on a specific date.

$$\mathrm{SGV}_i = \sum_{j=1}^{n_i} |\mathrm{N}_{ij}| \Big/ n_i,$$

where i is a day between January 1, 2020 and December 7, 2021, N_{ij} is a set of mutations identified from the complete genome j on the date of i, $|\mathrm{N}_{ij}|$ is the number of the mutations in N_{ij} and n_i is the number of the complete genomes collected on the date of i.

4.3 The pace of evolution evaluated by mutation change rate in SGV

To simulate how VOC competes with other variants, we need to model how the virus controls the mutational process by its mutation speed change. We defined the pace of evolution as a weekly change in the SGV indices. Specifically, we used derivative and second derivative to identify the accelerated and decreased paces. We denoted the pace of the evolution as $y' = f'(x) = \frac{dy}{dx} \sim \frac{\Delta y}{\Delta x} \sim \frac{\Delta \mathrm{SGV}}{\Delta x}$, and the peaks and valleys as $x|y'' = f''(x) = \frac{dy'}{dx} \sim \frac{\Delta y'}{\Delta x} = 0$, where x is the date, y is the SGV index and f is the regression function of y. To normalize the pace of evolution, we modified y' as

$$y'(x_{i+1}) \sim g(x_{i+1}) = \mathrm{pace}_{i+1} = \frac{\frac{\mathrm{SGV}_{i+1}-\mathrm{SGV}_i}{\mathrm{SGV}_i}}{\Delta x} \sim \frac{\mathrm{SGV}_{i+1}-\mathrm{SGV}_i}{\mathrm{SGV}_i} \text{ per week.}$$

Accordingly, the change in the pace of evolution was defined as

$$\mathrm{Pace}_{\mathrm{change}} = g'(x).$$

The peaks and valleys ($x|g'(x) = 0$) identified stages. Notably, the value of $g(x)$ changed within 1% per week was ignored in identifying stages.

4.4 Stages

Evolutionary stages were essential to VOC-alarm. Stages provided critical time periods to distinctly analyze the competition process, in which a VOC emerged from a small or a tiny population, and the global outbreak process, in which the emerged VOC caused a COVID surge. Stages were defined by the changes in the estimated pace of evolution. We defined a stage as a period in which the pace of evolution was either accelerated or decelerated. Mathematically,

$$x|g'(x) < 0 \text{ or } x|g'(x) > 0.$$

Accordingly, we classified the days from January 1, 2020 to December 7, 2021 into seven stages by the six dates ($x|g'(x) = 0$), that was, October 23, 2020, February 7, 2021, March 29, 2021, May 31, 2021, August 30, 2021 and October 20, 2021. In addition, we also predicted the future trend of the pace of evolution to March 10, 2022 (Fig. 5A), in which Stages VII and VIII were defined by the date of January 19, 2022 that satisfied $x|g'(x) = 0$ in our simulated data.

Stages I, III, V and VII showed accelerated paces of evolution ($g'(x) > 0$), but Stages II, IV, VI and VIII had decelerated paces ($g'(x) < 0$). Specifically, Stage I was from January 1, 2020 to October 23, 2020, Stage II was from October 24, 2020 to February 7, 2021; Stage III was from February 8, 2021 to March 29, 2021; Stage IV was from March 30, 2021 to May 31, 2021 and Stage V was from June 1, 2021 to September 28, 2021 (Fig. 1A). Using the latest GISAID data as of December 7, 2021, we modified Stage V as June 1, 2021–August 30, 2021 and added a new stage VI as August 31, 2021–October 20, 2021 and VII as October 21, 2021–December 7, 2021. Using our simulated future trend data, we extended Stage VII to January 19, 2022 and added a new stage VIII from January 9, 2021 to March 10, 2022. Specifically, we predicted VOCs, Alpha, Delta, Delta plus and Omicron in the Stages I, III, V and VII.

4.5 Mutational entropy

A difficult in understanding a VOC was from quantifying its adaptiveness to selective pressures from diagnostics, treatments and/or vaccines (Choi *et al.*, 2021; Collier *et al.*, 2021; Lopez Bernal *et al.*, 2021; Payne *et al.*, 2021; Sheikh *et al.*, 2021; Wilder-Smith and Mulholland, 2021). To accomplish this goal, we used the concept of entropy from the information theory (Ghanchi *et al.*, 2021; Tomaszewski *et al.*, 2020). Our defined mutational entropy for analyzing the competition among variants is distinct from the currently used strategy by phylogenetic analysis or those specific marker

mutations (Ascoli, 2021; Harvey *et al.*, 2021; Muecksch *et al.*, 2021; Wang *et al.*, 2021). Instead of comparing the mutated loci in the genomes, we considered the change of the mutation numbers across the globe for a specific lineage in a specific time period. Thus, this metric became feasible to simulate the competition among variants and predict VOCs by the lineages with the strongest competitive capabilities.

Within a time period (e.g. Stage I), we collected the variants $\{1, 2, \ldots, n^l\}$ from the lineage l across the days in this time period. For each variant j ($j = 1, 2, \ldots, n^l$), we defined N_{jl} is a set of mutations found in the complete genome of the variant j and $|N_{jl}|$ is the number of the mutations in N_{jl}. For all variants $\{1, 2, \ldots, n^l\}$ in the lineage l, we derived the unique mutation numbers from all mutation numbers $\{|N_{1l}|, |N_{2l}|, \ldots, |N_{n^l,l}|\}$ and denoted these unique mutation numbers as M_l. For each unique mutation number M_{kl} ($k = 1, 2, \ldots, u^l$), we calculated its frequency Q_{kl} ($k = 1, 2, \ldots, u^l$) and defined its probability $P_{kl} = \frac{Q_{kl}}{u^l}$ ($k = 1, 2, \ldots, u^l$). Thus, we defined the mutational entropy for this lineage within this time period as

$$\text{Entropy}_l = -\sum_{k=1}^{u_l} P_{kl} \log P_{kl}.$$

4.6 Sequence prevalence

Competition among variants not only determined VOCs but also their population size changes. It was well known that an emerged VOC tended to compete with its precedents, which caused the population size decrease of its precedents (Elbe and Buckland-Merrett, 2017; Vohringer *et al.*, 2021). We described the population size percentage at a specific time range as prevalence. Generally, we used the sequence number in the GISAID database to calculate prevalence. Specifically, using the GISAID database, if one lineage l included k_l sampled sequences ($l \in \{1, 2, \ldots, n_l\}$), its prevalence was defined as

$$\text{Prevalence}_l = \frac{k_l \times 100}{\sum_{i=1}^{n_l} k_i}\%.$$

The prevalence of this lineage for a specific date or time range could be defined by the sampled sequences on this date or within this time range.

4.7 VOC-alarm

VOC-alarm was a software for real-time monitoring of stages, competition among variants, VOCs and potential future COVID surges. Our defined Stages and calculated mutational entropy values were the key to predicting a VOC by its competition with others. The Stages I, III, V and VII, showing accelerated paces of evolution, predicted Alpha/B.1.1.7 (Chemaitelly *et al.*, 2021; Payne *et al.*, 2021; Washington *et al.*, 2021), Delta/AY.4 (Celik and Tallei, 2022), Delta/B.1.617.2 (Lopez Bernal *et al.*, 2021; Pung *et al.*, 2021; Sonabend *et al.*, 2021), Delta plus/AY.4.2 (Angeletti *et al.*, 2021) and Omicron/B.1.1.529 (Chen *et al.*, 2021) as VOCs, respectively.

Four criteria were used for identifying lineages/clades for predicting a potential VOC (Supplementary Fig. S6):

i. Stages for predicting VOCs should illustrate an accelerated pace of evolution ($g'(x) > 0$), e.g. Stages I, III, V and VII;
ii. In these stages, VOCs generally emerged from the highly mutated variants (generally, 95 percentiles of the mutation numbers were used as the threshold);
iii. In these stages, the precedents of the emerging VOCs showed a significant decrease in population size;
iv. In these stages, the emerging variants themselves grew in population size.

To ensure these four criteria and apply mutational entropy concept, we developed a flowchart for VOC-alarm, as shown in Supplementary Figure S6. We summarized this flowchart into the following four steps:

1. Identifying a stage with the accelerated pace of evolution ($x|g'(x) > 0$). This step enabled the prediction of the earliest emerging dates for VOCs ($x|g'(x) = 0$) in these stages. The earliest emerging variants within a small population were generally associated with the accelerated pace of evolution.
2. In an identified stage, the clade with the most significant decreased population size, evaluated by an ANOVA test, was used for predicting a VOC. This was because of the competition from the VOC and its precedents.
3. Variants with relatively high mutation numbers and increased population sizes were selected from the identified clades/lineages (in Step 2). We used a threshold for mutation numbers, i.e. 95 percentiles, and a population growth rate as 1 for selecting lineages as the candidates for the VOCs. Generally, we divided the days in this stage into two periods by its middle time point and calculated the population growth rate as the ratio of the sampled sequence numbers within the later period (late stage) and the earlier period (early stage).
4. The lineage(s) ranked at the top by mutational entropy was predicted as a VOC. This was determined by the strong competitive capabilities of VOCs. A relatively high mutational entropy value suggested a high potential to be a VOC in the competition with other variants.

In this work, we identified a threshold from lineage B.1.1, $\text{Entropy}_{B.1.1} = 4.25$, for identifying VOCs. We have applied this threshold to predict Alpha/B.1.1.7 (Chemaitelly *et al.*, 2021; Payne *et al.*, 2021; Washington *et al.*, 2021), Delta/AY.4 (26), Delta/B.1.617.2 (Lopez Bernal *et al.*, 2021; Pung *et al.*, 2021; Sonabend *et al.*, 2021), Delta plus/AY.4.2 (Angeletti *et al.*, 2021) and Omicron/B.1.1.529 (Chen *et al.*, 2021).

4.8 The pandemic epidemiological data

Despite that VOC-alarm did not use the pandemic epidemiological data, we have considered the epidemiological data of new cases and new deaths to verify the COVID surges caused by VOCs. We used the epidemiological data sources from the WHO, Our world in Data (owid: https://ourworldindata.org/), COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University (Dong *et al.*, 2020), data from The New York Times (https://github.com/nytimes/covid-19-data), based on reports from state and local health agencies and data from Coronavirus (COVID-19) in the UK (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/).

4.9 Dates of designation of VOCs and VOIs from the WHO and the US Centers for Disease Control and Prevention

VOCs or VOIs have been designated by the WHO and the US Centers for Disease Control and Prevention. The information of designation, e.g. Pango lineage name, the GISAID clade name and the designated dates, were included in Supplementary Table S14.

4.10 Statistical analyses

All statistical analyses and visualization were conducted using R (version 4.4.1) or Python (version 3.8.3). Barplot, boxplot and scatter plot were constructed using R package ggplot2 (version 3.3.5); stream plots were implemented using R package ggstream (version 0.1.0); the R package for entropy (version 1.3.0) with the ML method, was used to calculate entropy scores; ROC was calculated with the Python package, sklearn. The LOESS method was used for regression of the pace of evolution. Span of 0.5 was used for the mutation data as of September 28, 2021 and 0.25 was used for the

mutation data as of December 7, 2021. The regression on the future trend for Omicron was based on the trend from Delta plus. *P*-value <0.05 from two-sided Student's *t*-test or one-way ANOVA test was considered significant.

Acknowledgements

We acknowledge all the researchers who openly shared their genomic data on GISAID.

Author contribution

G.J. was responsible for conception and design of the study. G.J. designed the computational analysis software. H.Z., K.H. and C.G. were responsible for software development. H.Z., V.M., U.T., and G.J. took part in data analyses. H.Z., V.M., U.T., Y.L. and G.J. wrote the manuscript.

Funding

This work was supported by a start-up fund from Wake Forest University Health Sciences. We also acknowledge assistance of the Wake Forest Baptist Comprehensive Cancer Center Bioinformatics Shared Resource, supported by [P30CA012197]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute.

Conflict of Interest: none declared.

Data and software availability

All datasets used in our study were listed in Supplementary Table S15. We downloaded the SARS-CoV-2 mutation data from GISAID (Elbe and Buckland-Merrett, 2017). Due to the restriction on the availability of the raw data and meta data from GISAID, requests for these data should be submitted to GISAID directly. Processed data related to our results were included in the Supplementary Tables.

References

Aljindan,R.Y. *et al.* (2021) Investigation of nonsynonymous mutations in the spike protein of SARS-CoV-2 and its interaction with the ACE2 receptor by molecular docking and MM/GBSA approach. *Comput. Biol. Med.*, **135**, 104654.

Angeletti,S. *et al.* (2021) SARS-CoV-2 AY.4.2 variant circulating in Italy: genomic preliminary insight. *J. Med. Virol.*, **94**, 1689–1692.

Arbeitman,C.R. *et al.* (2021) The SARS-CoV-2 spike protein is vulnerable to moderate electric fields. *Nat. Commun.*, **12**, 5407.

Arora,P. *et al.* (2022) No evidence for increased cell entry or antibody evasion by Delta sublineage AY.4.2. *Cell. Mol. Immunol.*, **19**, 449–452.

Ascoli,C.A. (2021) Could mutations of SARS-CoV-2 suppress diagnostic detection? *Nat. Biotechnol.*, **39**, 274–275.

Benvenuto,D. *et al.* (2020) The global spread of 2019-nCoV: a molecular evolutionary analysis. *Pathog. Glob. Health*, **114**, 64–67.

Celik,I. and Tallei,T. (2022) A computational comparative analysis of the binding mechanism of molnupiravir's active metabolite to RNA-dependent RNA polymerase of wild-type and Delta subvariant AY.4 of SARS-CoV-2. *J. Cell. Biochem.*, **123**, 807–818.

Chakraborty,C. *et al.* (2021) D614G mutation eventuates in all VOI and VOC in SARS-CoV-2: is it part of the positive selection pioneered by Darwin? *Mol. Ther. Nucleic Acids*, **26**, 237–241.

Chemaitelly,H. *et al.* (2021) mRNA-1273 COVID-19 vaccine effectiveness against the B.1.1.7 and B.1.351 variants and severe COVID-19 disease in Qatar. *Nat. Med.*, **27**, 1614–1621.

Chen,J. *et al.* (2021) Omicron Variant (B.1.1.529): Infectivity, Vaccine Breakthrough, and Antibody Resistance. *J. Chem. Inf. Model*, **62**, 412–422.

Choi,A. *et al.* (2021) Safety and immunogenicity of SARS-CoV-2 variant mRNA vaccine boosters in healthy adults: an interim analysis. *Nat. Med.*, **27**, 2025–2031.

Collier,D.A. *et al.*; CITIID-NIHR BioResource COVID-19 Collaboration. (2021) Age-related immune response heterogeneity to SARS-CoV-2 vaccine BNT162b2. *Nature*, **596**, 417–422.

Davies,N.G. *et al.*; CMMID COVID-19 Working Group. (2021) Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, **372**, eabg3055.

Del Rio,C. *et al.* (2021) Confronting the Delta variant of SARS-CoV-2, summer 2021. *JAMA*, **326**, 1001–1002.

Dhar,M.S. *et al.*; The Indian SARS-CoV-2 Genomics Consortium (INSACOG). (2021) Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. *Science*, **374**, 995–999.

Dong,E. *et al.* (2020) An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.*, **20**, 533–534.

Elbe,S. and Buckland-Merrett,G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.*, **1**, 33–46.

Fan,Z. *et al.* (2021) Entropy-driven amplified electrochemiluminescence biosensor for RdRp gene of SARS-CoV-2 detection with self-assembled DNA tetrahedron scaffolds. *Biosens. Bioelectron.*, **178**, 113015.

Fan,Z. *et al.* (2022) Rational engineering the DNA tetrahedrons of dual wavelength ratiometric electrochemiluminescence biosensor for high efficient detection of SARS-CoV-2 RdRp gene by using entropy-driven and bipedal DNA walker amplification strategy. *Chem. Eng. J.*, **427**, 131686.

Fariselli,P. *et al.* (2021) DNA sequence symmetries from randomness: the origin of the Chargaff's second parity rule. *Brief. Bioinform.*, **22**, 2172–2181.

Ghanchi,N.K. *et al.* (2021) Higher entropy observed in SARS-CoV-2 genomes from the first COVID-19 wave in Pakistan. *PLoS One*, **16**, e0256451.

Graham,F. (2020) Daily briefing: pangolins return to a region where they were once extinct. *Nature*.

Harvey,W.T. *et al.*; COVID-19 Genomics UK (COG-UK) Consortium. (2021) SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.*, **19**, 409–424.

Hung,I.F. *et al.* (2020) Triple combination of interferon beta-1b, lopinavir-ritonavir, and ribavirin in the treatment of patients admitted to hospital with COVID-19: an open-label, randomised, phase 2 trial. *Lancet*, **395**, 1695–1704.

Jentsch,P.C. *et al.* (2021) Prioritising COVID-19 vaccination in changing social and epidemiological landscapes: a mathematical modelling study. *Lancet Infect. Dis.*, **21**, 1097–1106.

Kannan,S.R. *et al.* (2021) Evolutionary analysis of the Delta and Delta plus variants of the SARS-CoV-2 viruses. *J. Autoimmun.*, **124**, 102715.

Kontis,V. *et al.* (2020) Magnitude, demographics and dynamics of the effect of the first wave of the COVID-19 pandemic on all-cause mortality in 21 industrialized countries. *Nat. Med.*, **26**, 1919–1928.

Lopez Bernal,J. *et al.* (2021) Effectiveness of covid-19 vaccines against the B.1.617.2 (Delta) variant. *N. Engl. J. Med.*, **385**, 585–594.

McBroome,J. *et al.* (2021) A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. *Mol. Biol. Evol.*, **38**, 5819–5824.

Muecksch,F. *et al.* (2021) Affinity maturation of SARS-CoV-2 neutralizing antibodies confers potency, breadth, and resilience to viral escape mutations. *Immunity*, **54**, 1853–1868.e1857.

Mukherjee,S. *et al.* (2013) Cell responses only partially shape cell-to-cell variations in protein abundances in *Escherichia coli* chemotaxis. *Proc. Natl. Acad. Sci. USA*, **110**, 18531–18536.

Narykov,O. *et al.* (2021) Computational protein modeling and the next viral pandemic. *Nat. Methods*, **18**, 444–445.

Payne,R.P. *et al.*; PITCH Consortium. (2021) Immunogenicity of standard and extended dosing intervals of BNT162b2 mRNA vaccine. *Cell*, **184**, 5699–5714.e5611.

Pung,R. *et al.*; CMMID COVID-19 Working Group. (2021) Serial intervals in SARS-CoV-2 B.1.617.2 variant cases. *Lancet*, **398**, 837–838.

Saito,A. *et al.* (2021) Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. *Nature*, **602**, 300–306.

Saunders,N. *et al.* (2022) Fusogenicity and neutralization sensitivity of the SARS-CoV-2 Delta sublineage AY.4.2. *EBioMedicine*, **77**, 103934.

Sheikh,A. *et al.*; Public Health Scotland and the EAVE II Collaborators. (2021) SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. *Lancet* **397**, 2461–2462.

Sonabend,R. *et al.* (2021) Non-pharmaceutical interventions, vaccination, and the SARS-CoV-2 delta variant in England: a mathematical modelling study. *Lancet*, **398**, 1825–1835.

Thompson,C.N. *et al.*; PhD1. (2021) Rapid emergence and epidemiologic characteristics of the SARS-CoV-2 B.1.526 Variant - New York city, New York, January 1-April 5, 2021. *MMWR Morb. Mortal Wkly. Rep.*, **70**, 712–716.

Tomaszewski,T. *et al.* (2020) New pathways of mutational change in SARS-CoV-2 proteomes involve regions of intrinsic disorder important for virus replication and release. *Evol. Bioinform. Online*, **16**, 1176934320965149.

Vohringer,H.S. *et al.* (2021) Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature*, **600**, 506–511.

Wang,L. *et al.* (2021) Ultrapotent antibodies against diverse and highly transmissible SARS-CoV-2 variants. *Science*, **373**, eabh1766.

Wang,R. *et al.* (2021) Analysis of SARS-CoV-2 variant mutations reveals neutralization escape mechanisms and the ability to use ACE2 receptors from additional species. *Immunity*, **54**, 1611–1621 e1615.

Washington,N.L. *et al.* (2021) Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *Cell*, **184**, 2587–2594.e2587.

Wilder-Smith,A. and Mulholland,K. (2021) Effectiveness of an inactivated SARS-CoV-2 vaccine. *N. Engl. J. Med.*, **385**, 946–948.

Zhang,H. *et al.* (2020) Ethics committee reviews of applications for research studies at 1 hospital in China during the 2019 novel coronavirus epidemic. *JAMA*, **323**, 1844–1846.