Published in final edited form as:
J Chem Inf Model. 2023 January 09; 63(1): 335–342. doi:10.1021/acs.jcim.2c01352.

Emerging dominant SARS-CoV-2 variants

Jiahui Chen[1], **Rui Wang**[1], **Yuta Hozumi**[1], **Gengzhuo Liu**[1], **Yuchi Qiu**[1], **Xiaoqi Wei**[1], **Guo-Wei Wei**[1,2,3,*]

[1]Department of Mathematics, Michigan State University, MI 48824, USA.

[2]Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA.

[3]Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA.

Abstract

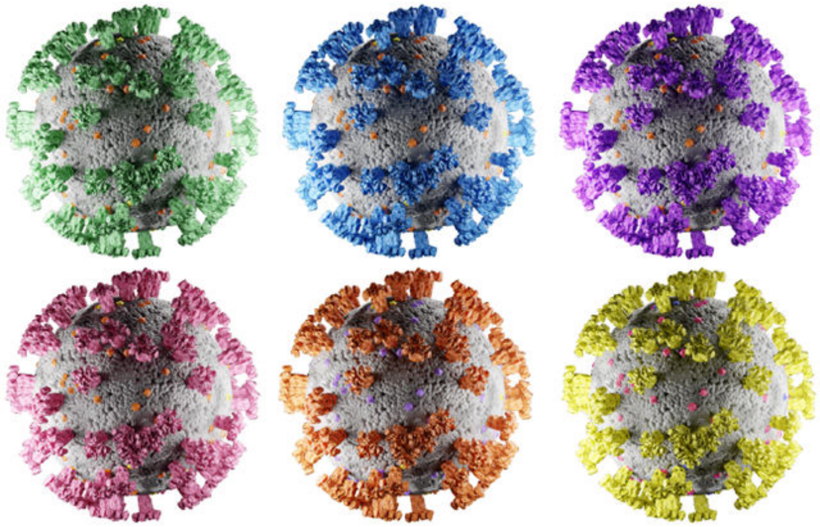
Accurate and reliable forecasting of emerging dominant severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants enables policymakers and vaccine makers to get prepared for future waves of infections. The last three waves of SARS-CoV-2 infections caused by dominant variants Omicron (BA.1), BA.2, and BA.4/BA.5 were accurately foretold by our artificial intelligence (AI) models built with biophysics, genotyping of viral genomes, experimental data, algebraic topology, and deep learning. Based on newly available experimental data, we analyzed the impacts of all possible viral spike (S) protein receptor-binding domain (RBD) mutations on the SARS-CoV-2 infectivity. Our analysis sheds light on viral evolutionary mechanisms, i.e., natural selection through infectivity strengthening and antibody resistance. We forecast that BP.1, BL*, BA.2.75*, BQ.1*, and particularly, BN.1*, have high potential to become new dominant variants to drive the next surge. Our key projection about these variants dominance made on Oct. 18, 2022 (see arXiv:2210.09485) became reality in late November 2022.

Graphical Abstract

[*] Corresponding author. weig@msu.edu.

Supporting information
The Supporting information (Supporting_Material.zip) includes the predicted BFE changes for all possible RBD mutations based on BA.1 and BA.2 structures, PDB files (6M0J, 7T9L, 7XB0) and a README file for the links of external datasets.

Data and model availability
The world's SARS-CoV-2 SNP data are available at Mutation Tracker (https://users.math.msu.edu/users/weig/SARS-CoV-2_Mutation_Tracker.html). The TopNetTree model is available at the repository (https://github.com/WeilabMSU/TopNetmAb).



1 Introduction

In the past two years, the coronavirus disease-2019 (COVID-19) pandemic was fueled by the spread of a few dominant variants of severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2), as shown in Figure 1. Specifically, the Alpha and Beta variants contributed to a peak of infections and deaths from October 2020 to January 2021. The Gamma variant caused another peak of infections and deaths in April and May 2021. The Delta variant led to the third wave of COVID-19 infections and deaths around August 2021. The Omicron (B.1.1.529), which was extraordinary in its infectivity, vaccine breakthrough, and antibody resistance, created a huge spike in the world's daily infection record in December 2021 and January 2022. Omicron BA.2 subvariant rapidly replaced the original Omicron (i.e., BA.1) in March 2022. Around July 2022, Omicron subvariants BA.4 and BA.5 took over BA.2 and became the new dominant SARS-CoV-2 variant. These variant-driven waves of infections are also associated with spikes in deaths and have given rise to tremendous economic loss. A life-and-death question is: what will be future dominant variants?

Forecasting and surveillance of emerging SARS-CoV-2 variants are some of the most challenging tasks of our time. Among about half a million SARS-CoV-2/COVID-19 related publications recorded in Google Scholar, few accurately foretold the emerging SARS-CoV-2 variants. Accurate and reliable forecasting of emerging SARS-CoV-2 variants enables policymakers and vaccine makers to plan, leading to enormous social, economic, and health benefits. To foretell future variants, one must have the full understanding of the mechanisms of viral evolution, the mechanisms of viral mutations, and the relationship between viral evolution and viral mutation.

Future variants are created through the SARS-CoV-2 evolution, in which is a SARS-CoV-2 evolves through changes in its RNA at molecular scale to gain fitness over its counterparts at the host population scale. At the molecular scale, most mutations occur randomly. Indeed, random genetic drift is a major mechanism of mutations, resulting in errors in various biological processes, such as replication, transcription, and translation. Additionally, virus-

virus intra-organismic recombination can alter SASR-CoV-2 genes, which has a stochastic nature too. However, SARS-CoV-2 has a genetic proofreading mechanism facilitated by the synergistic interactions between RNA-dependent RNA polymerase and non-structure proteins 14 (NSP14) [2, 3]. At the organismic scale, inter-organismic recombination happens but the resulting variants may not be clinically significant. In contrast, host editing of virus genes is known to be a significant mechanism for SARS-CoV-2 mutations [4]. At the population scale, mutations occurring at molecular and organismic scales are regulated, i.e., enhanced and/or suppressed via natural selection, giving rise to SARS-CoV-2 variants with increased fitness [5]. Therefore, natural selection is the fundamental driven force for viral evolution.

It remains to understand what controls the natural selection of SARS-CoV-2. The mechanism of SARS-CoV-2 evolution was elusive at the beginning of the COVID-19 pandemic. Indeed, the life cycle of SARS-CoV-2 is extremely sophisticated, involving the viral entry of host cells, the release of the viral genome, the synthesis of viral NSPs, RNA replication, the transcription, translation, and synthesis of viral structural proteins, and the packing, assembly, and release of new viruses [6]. The SARS-CoV-2 mutations occur nearly randomly on all of its 29 genes, as shown in Figure 2. Nonetheless, in early 2020, we hypothesized that SARS-CoV-2 natural selection is controlled through infectivity-strengthening mutations [5], which primarily occur at the viral spike (S) protein receptor-binding domain (RBD) that binds with host angiotensin-converting enzyme 2 (ACE2) to facilitate the viral cell entry [7-11]. Our hypothesis was initially supported by our genotyping of 15,140 SARS-CoV-2 genomes extracted from patients. We demonstrated that among 89 unique RBD mutations, the observed frequencies of infectivity-strengthening mutations outpace those of infectivity-weakening ones in their time evolution. Our infectivity-strengthening mechanism of natural selection was proven beyond doubt in April 2021, with 506,768 SARS-CoV-2 genomes isolated from patients [12].

However, we found that not all of the most observable RBD mutations strengthen viral infectivity [13]. This exception took place in the middle and late 2021 when a good portion of the population in many developed countries was vaccinated. By the genotyping of 2,298,349 complete SARS-CoV-2 genomes, we discovered vaccination-induced antibody-resistant mutations, which make the virus less infectious [13]. This discovery leads to a complementary mechanism of natural selection, namely antibody-resistant mutations. In other words, viral evolution also favors RBD mutations in a population that enable the virus to escape antibody protection generated from vaccination or infection.

The Omicron variant was the first example that was induced by both infectivity strengthening and antibody resistance mechanisms [13]. It has 32 mutations on the S protein, the main antigenic target of antibodies [14]. Among them, 15 are on the Omicron RBD, leading to a dramatic increase in SARS-CoV-2 infectivity, vaccine breakthrough, and antibody resistance [15]. The World Health Organization (WHO) declared Omicron as a variant of concern (VOC) on November 26, 2021. On December 1, 2021, when there were no experimental data available, we announced our topological artificial intelligence (AI) predictions based on the genotyping of viral genomes, biophysics, experimental data of protein-protein interactions, algebraic topology, and deep learning [16]. We predicted that

Omicron is about 2.8 times as infectious as the Delta and has nearly 90% likelihood to escape vaccines, which would compromise essentially all of existing monoclonal antibody (mAb) therapies from Eli Lilly, Regeneron, AstraZeneca, etc. These predictions were subsequently confirmed by experiments [**hoffmann2021omicron**, 14, 17-20]. On February 10, our topological AI model foretold the taking over of Omicron BA.1 by Omicron subvariant BA.2 [21]. The WHO declared BA.2's dominance on March 26, 2022. On May 1, 2022, our topological AI model projected the incoming dominance of BA.4 and BA.5 [22], which became reality in late June 2022. Currently, BA.5 is still the world's dominant variant. Therefore, our topological AI model has been offering unusually accurate two-month forecasts of emerging dominant variants.

The COVID fatigue and the worldwide relaxation of COVID-19 prevention measures have given the virus enormous new opportunities to spread in world populations, which enables the virus to further evolve. Additionally, the newly generated Omicron subvariant RBD structures leave abundant room for the virus to further optimize its binding with the ACE2 and disrupt existing antibodies, resulting in a large number of emerging subvariants. It is of paramount importance to analyze their growth potentials in the world's populations and alert future dominant variants.

This work analyzes SARS-CoV-2 evolutionary trends. We predict the SARS-CoV-2 S protein RBD mutation-induced binding free energy (BFE) changes of RBD-human ACE2 complexes at all RBD residues. Such changes are employed to forecast Omicron subvariants' growth potentials and chances of becoming future dominant variants. Topological AI models are built from newly available deep mutational screening data and Omicron BA.1 and BA.2 three-dimensional (3D) structures. Our studies are assisted with the genotyping of over three million SARS-CoV-2 genomes extracted from patients and the evolutionary pattern of viral lineages among infections in the United States. Our key projection of emerging variants incoming dominance made in Oct 18, 2022 [23] had become reality in late November 2022.

2 Results

We carry out single nucleotide polymorphism (SNP) calling for 3,616,783 million complete genomes extracted from patients. All unique mutations and their observed frequencies are illustrated in Fig. 2. Our interactive website, Mutation Tracker, also provides detailed records of mutations for download. On average, each nucleic acid site has one mutation. Overall, mutations occur essentially randomly at all 29,903 bases. Therefore, simple SNP calling and genotyping does not offer any direct evidence for SARS-CoV-2 variants as discussed earlier. More specific analysis of the RBD mutations is used for the forecasting of future dominant variants.

We collect emerging Omicron sublineages and compare them with previous VOCs. In Fig. 3, we preset the annotation tree plot of recently occurred Omicron subvariants. Mutations from parent generations to children are marked on edges as well as binding free energy (BFE) changes (kcal/mol) induced by the corresponding mutations. As many as 106 Omicron subvariants and their relationships are delineated in the plot.

We use the notation “*” to represent the lineage and its sublineage. For instance, BA.2* represents BA.2 and all its sublineages in Figure 4. Figure 4a and b show the 3D structures of RBD binding to human ACE2. Figure 4a includes the RBD mutations of previous VOCs and Omicron subvariants BA.1, BA.2, BA.3, BA.4, and BA.5, while Figure 4b shows mutations of the subvariants of Omicron BA.2 and BA.5. Lineages originated from BA.2 are marked in red type of colors. Subvariants originated from BA.5 are labeled in green type of colors. In Figure 4c, the BFE changes of previous VOCs, BA.1 and BA.2 are calculated as the accumulation of single mutations according to the original structure (PDB: 6M0J [25]). The BFE changes of BA.1.1 is calculated based on the BA.1 RBD-ACE2 structure (PDB: 7T9L [26]). For the sublineages of BA.2, as well as BA.3, BA.4 and BA.5, their BFE changes are calculated based on the BA.2 structure (PDB: 7XB0 [24]). In Figures 4d and e, the BFE changes of the BA.2, BA.4, and BA.5’s sublineages are presented.

In Figure 4c, the variants prior to the Omicron are presented in light blue including previous VOCs, BA.1, BA.1.1, BA.2, BA.3, and BA.4. In Figure 4d, there are three main clades, one from BA.2, one from BA.4, and the other from BA.5. Firstly, three mutations from BA.2 to BA.5 are L452R, F486V, and R493Q, which make BA.5 two-fold as infectious as BA.2. This explains why BA.5 replaced BA.2 as a new dominant variant in late June 2022. Among the BA.2 sublineages, BP.1, BA.2.10.4, BA.2.3.*, BA.2.75.*, BL.1.*, BR.*, BN.1.*, and CB.1 were predicted to have BFE changes greater than 4.0 kcal/mol. These three sublineages together with BA.2.10.4 and BA.2.75.2 have higher BFE changes and are more infectious than BA.4 and BA.5. As for BA.4 and BA.5 sublineages, BA.4.6 is more infectious than BA.4 and BA.5 and has potential to become a dominant variant and its sublineage BA.4.6.3 has a BFE change greater than 4.0 kcal/mol. Among the sublineages of BA.5, BQ.1.1 has the highest potential to replace the spreading of BA.5 as its BFE change is greater than 4.0 kcal/mol, while some of BA.5’s sublineages BF.7, BQ.1, and BE.1.2 have larger BFE changes than that of BA.5. Based on this analysis shown in Figure 4d and e, we forecast that BP.1, BL*, BA.2.75*, BQ.1*, and particularly, BN.1*, have high potentials to become new dominant variants.

Figure 5 shows the heatmaps of predicted mutation-induced BFE change predictions of BA.1 (top panel) and BA.2 (bottom panel) variants. We plot those RBD residues that have at least one mutation-induced BFE change greater than 0.1 kcal/mol, which gives rise to 89 residues in the plots. In other words, we keep mutations that will lead to more infectious variants. The deep blue color indicates infectivity-strengthening mutations. Deep red color shows infectivity-weakening mutations. It is seen from Figure 5 that most mutations will weaken the binding between RBD and ACE2 for BA.1 and BA.2. However, such mutations, once occurred, will have little chance of becoming clinically significant due the natural selection. Figure 5 indicates that both BA.1 and BA.2 are highly infectivity-optimized variants. They just leave a few residues to be further optimized. Obviously, for both B.1 and BA.2, many mutations on residue sites R439, Y453, and N417 will most likely lead to more infectious new variants. For BA.2, surveillance is also required for residue sites N504 and R403.

Compared with BA.2, BA.5 has three additional mutations, i.e., L452R, F486V, and R493Q. Among them, R493Q makes BA.5 significantly more infectious as shown in Figure 5. This reverse mutation (original residue is glutamine) occurs in many other lineages showing in Figure 4c, namely, BA.2.10.4, BA.2.75*, BA.4*, BA.5, BF.7, BQ.1* and BE.1.2. In addition to R493Q, BA.2.75* and BQ.1.1 in Figure 4 share the mutation N460K with the BFE change 0.267 kcal/mol. This indicates that more infectious variants will emerge with multiple infectivity-strengthening mutations. Overall, comparing the two heatmaps in Figure 5, it is easy to note that BA.2 has more positive BFE changes, which makes future BA.2 sublineages more competitive than future BA.1 sublineages in terms of infectivity.

The top panel of Figure 5 explains why BA.2 is more infectious than BA.1. BA.2 shares 12 of its RBD mutations with BA.1, except for six mutations, i.e., L371F, T376A, D405N, R408S, S446G, and S496G. These residue sites are marked with red in both panels of Figure 5. Among these mutations, L371F, T376A, D405N, and R408S induced minor BFE changes as shown in the top panel of Figure 5. However, S446G and S496G render BA.2 significantly more infectious than BA.1.

3 Discussion

Figure 6 presents the evolution pattern of weekly viral lineage distribution among infections in the United States from 06/26/2022 to 11/26/2022 from CDC website[27]. Each lineage is illustrated by aggregating its sublineages to except for its sublineage is also listed. Note that BA.2.75 sublineages except BA.2.12.1, BA.2.75, BA.2.75.2, BN.1, XBB and their sublineages are aggregated to BA.2.75, which means lineages BA.2.10.4 in Figure 4 belong to this category. It is interesting to note that there is high consistence between Figures 4 and 6. Specifically, all the emerging variants listed in Figure 6 have relatively high BFE changes as depicted in Figure 4.

It is also interesting to note from Figure 6 that the relative populations of BA.2.12.1, BA.4, and BA.5 are shrinking during this period. BA.5 slightly expanded at the beginning and took a portion of BA.2.12.1’s population. BA.4.6 is a sublineage of BA.4, while BF.7, BQ.1, and BQ.1.1 are the sublineages of BA.5. Their relative populations are increasing. BQ.1.1 has a faster growth rate than BQ.1 and BF.7, which indicates that the predicted BFE change of BQ.1.1 is the highest among the sublineages of BA.5. As shown in Figure 4, BA.2.75, and BQ.1.1 have higher potentials to become future dominant variants.

In our earlier predictions of Omicron [15] and BA.2 [21], we utilized nearly 200 antibody-RBD complexes to analyze the impact of antibody resistance. Such analysis is necessary because the Omicron variant involves a dramatic increase in the number of RBD mutations. For the most variants studied in the present work, there are only gradual changes in the number of new RBD mutations and thus the impact of antibody resistance on natural selection may be relatively small, particularly for the population that has not been exposed to Omicron and its subvariants.

While the BFE change-based prediction favors the variant with the highest BFE change, its dominance in the population is also determined by the viral transmission environment (i.e., vaccination, prevention measures, human interaction intensity, etc.) and temporal dynamics. Therefore, a variant with slightly lower BFE change might become a dominant variant over a short period, which is called kinetic reaction control in thermodynamics. In an idealized viral transmission environment, the variant with the highest BFE change would have an exponential advantage over other variants, according to the Boltzmann distribution, which is called thermodynamic reaction control.

4 Methods

4.1 Deep learning model

The model applied in this work is an updated version of the recently proposed machine learning model, TopLapNet, by integrating the SKEMPI 2.0 dataset [28] and deep mutational scanning datasets [29-32]. Briefly speaking, the TopLapNet model is a deep neural network model and implements biophysics and biochemistry descriptors, as well as mathematical descriptors based on algebraic topology [33-35] to predict the binding free energy (BFE) changes of protein-protein interactions (PPIs) induced by single mutations. A deep neural network maps sample features to an output layer where hidden layers in the network contain numerous neuron units and weights updated by backpropagation methods. The single neuron gets fully connected with the neurons in the following layers. For the model cross-validations, the Pearson correlation of 10-fold cross-validation is 0.864, and the root mean square error is 1.019 kcal/mol. As for predictions, the TopLapNet model is used to calculate all possible mutation impacts on RBD binding to ACE2 for the original virus (PDB: 6M0J [25]), BA.1 (PDB: 7T9L [26]), and BA.2 structures (PDB: 7XB0 [24]). Thus, previous VOCs' infectivities as well as that of BA.1 and BA.2 are calculated based on the original structure. The infectivity of BA.1.1 is calculated by accumulating BFE changes based on the BA.1 structure. The infectivities of all other sublineages presented in Figure 4 are calculated by the accumulations of BFE changes based on the BA.2 structure.

4.2 Feature generation

Feature generation methods decipher protein structures to extract their biophysics, biochemistry, and mathematical information. These methods use physical, chemical, and mathematical modeling of protein structures to provide suitable features for machine-learning algorithms. There are two types of features, i.e., residue-level ones and atom-level ones. Residue-level features are generated from secondary structures, which are provided by a position-specific scoring matrix (PSSM) in the form of conservation scores of each amino acid [36]. Atom-level features consider seven groups of atom types, including C, N, O, S, H, all heavy atoms, and all atoms. Surface areas, partial changes, atomic pairwise interactions, and electrostatics are assembled in an element-specific manner in terms of these seven groups. Moreover, the most important features from modelings are topological features and graph features generated by using persistent homology [**edelsbrunner2008persistent** , 33] and persistent Laplacian [35].

Persistent homology describes proteins by analogy to point cloud data. Atoms are regarded as vertices to build a simplicial complex, which is a collection of infinitely many simplicies such as nodes, edges, triangles, and tetrahedrons. The simplicies among atoms are defined by whether there is an overlap under a given influence domain or radius *r*. Filtration of this topological space is defined by varying the radius as a sequence of snapshots of each simplicial complex to extract more geometric and topological properties. Then, the Betti numbers on each snapshot are computed as descriptors of the number of connected components, cycles, and cavities in a protein structure. Persistent Laplacian (also known as persistent spectral graph [35, 37]) on the other hand unveil the homotopic shape evolution of a protein structure in filtration that the persistent homology cannot provide. It has been tested for its performance in mutation-induced PPI binding affinity change prediction [22, 37]. Persistent Laplacian applies the same scheme as persistent homology to construct simplicial complexes during filtration. However, persistent Laplacian calculates all eigenvalues of the combinatorial Laplacian with boundary operators on simplicial complexes. Our mathematical features consist of both topological invariants from persistent homology and spectral invariants from persistent Laplacian.

4.3 SNP calling and Mutation Tracker

For genotyping, SARS-CoV-2 complete genome sequences with high coverage and exact collection date were downloaded from the GISAID database [38] (https://www.gisaid.org/) as of September 30, 2022. Such sequences were aligned to the reference genome downloaded from GenBank (NC_045512.2)[39]. Next, we applied single nucleotide polymorphism (SNP) calling [40, 41] to measure the genetic variations between SARS-CoV-2 sequences through Cluster Omega with default parameters. The SNP calling can track differences between various SARS-CoV-2 sequences and the reference genome. By applying it, we decoded 29,290 unique single mutations from more than 3.6 million complete SARS-CoV-2 genomes. The detailed mutation information can be viewed at Mutation Tracker. Lastly, the Omicron sublineages analyzed in Figure 4 are selected from the SNP analysis and other web-servers [27, 42, 43].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

This work was supported in part by NIH grants R01GM126189 and R01AI164266, NSF grants DMS-2052983, DMS-1761320, and IIS-1900473, NASA grant 80NSSC21M0023, MSU Foundation, Bristol-Myers Squibb 65109, and Pfizer.

References

(1). Daily New Cases, https://www.worldometers.info/coronavirus/, Accessed: 2022-10-16.

(2). Sevajol M; Subissi L; Decroly E; Canard B; Imbert I Virus Research 2014, 194, 90–99. [PubMed: 25451065]

(3). Ferron F; Subissi L; De Morais ATS; Le NTT; Sevajol M; Gluais L; Decroly E; Vonrhein C; Bricogne G; Canard B; Imbert I Proceedings of the National Academy of Sciences 2018, 115, E162–E171.

(4). Wang R; Hozumi Y; Zheng Y-H; Yin C; Wei G-W Viruses 2020, 12, 1095. [PubMed: 32992592]

(5). Chen J; Wang R; Wang M; Wei G-W J. Mol. Biol 2020, 432, 5212–5226. [PubMed: 32710986]

(6). Trougakos IP; Stamatelopoulos K; Terpos E; Tsitsilonis OE; Aivalioti E; Paraskevis D; Kastritis E; Pavlakis GN; Dimopoulos MA Journal of Biomedical Science 2021, 28, 1–18. [PubMed: 33388061]

(7). Li W; Shi Z; Yu M; Ren W; Smith C; Epstein JH; Wang H; Crameri G; Hu Z; Zhang H; Zhang J; McEachern J; Field H; Daszak P; Eaton BT; Zhang S; Wang L-F Science 2005, 310, 676–679. [PubMed: 16195424]

(8). Qu X-X; Hao P; Song X-J; Jiang S-M; Liu Y-X; Wang P-G; Rao X; Song H-D; Wang S-Y; Zuo Y; Zheng A-H; Luo M; Wang H-L; Deng F; Wang H-Z; Hu Z-H; Ding M-X; Zhao G-P; Deng H-K Journal of Biological Chemistry 2005, 280, 29588–29595. [PubMed: 15980414]

(9). Song H-D; Tu C-C; Zhang G-W; Wang S-Y; Zheng K; Lei L-C; Chen Q-X; Gao Y-W; Zhou H-Q; Xiang H; Zheng H-J; Chern S-WW; Cheng F; Pan C-M; Xuan H; Chen S-J; Luo H-M; Zhou D-H; Liu Y-F; He J-F; Qin P-Z; Li L-H; Ren Y-Q; Liang W-J; Yu Y-D; Anderson L; Wang M; Xu R-H; Wu X-W; Zheng H-Y; Chen J-D; Liang G; Gao Y; Liao M; Fang L; Jiang L-Y; Li H; Chen F; Di B; He L-J; Lin J-Y; Tong S; Kong X; Du L; Hao P; Tang H; Bernini A; Yu X-J; Spiga O; Guo Z-M; Pan H-Y; He Wei-Zhong Manuguerra J-C; Fontanet A; Danchin A; Niccolai N; Li Y-X; Wu C-I; Zhao G-P Proceedings of the National Academy of Sciences 2005, 102, 2430–2435.

(10). Hoffmann M; Kleine-Weber H; Schroeder S; Krüger N; Herrler T; Erichsen S; Schiergens TS; Herrler G; Wu N-H; Nitsche A; Muller MA; Drosten C; Pohlmann S Cell 2020, 181, 271–280. [PubMed: 32142651]

(11). Walls AC; Park Y-J; Tortorici MA; Wall A; McGuire AT; Veesler D Cell 2020, 181, 281–292. [PubMed: 32155444]

(12). Wang R; Chen J; Gao K; Wei G-W Genomics 2021, 113, 2158–2170. [PubMed: 34004284]

(13). Wang R; Chen J; Wei G-W The Journal of Physical Chemistry Letters 2021, 12, 11850–11857. [PubMed: 34873910]

(14). Cele S; Jackson L; Khoury DS; Khan K; Moyo-Gwete T; Tegally H; San JE; Cromer D; Scheepers C; Amoako DG; Karim F; Bernstein M; Lustig G; Archary D; Ganga Y; Jule Z; Reedoy K; Hwa S-H; Giandhari J; Blackburn JM; Gosnel BI; Karim SSA; Hanekom W; NGS-SA; Team, C.-K.; von Gottberg A; Bhiman JN; Lessells RJ; Moosa M-YS; Davenport MP; Oliveira T. d.; Moore PL; Sigal A Nature 2021, 1–5.

(15). Chen J; Wang R; Gilby NB; Wei G-W J Chem Inf Model 2022, 62, 412–422. [PubMed: 34989238]

(16). Chen J; Wang R; Gilby N; Wei G arXiv preprint arXiv:2112.01318.

(17). Shuai H; Chan JF-W; Hu B; Chai Y; Yuen TT-T; Yin F; Huang X; Yoon C; Hu J-C; Liu H; Shi J; Liu Y; Zhu T; Zhang J; Hou Y; Wang Y; Lu L; Cai J-P; Zhang AJ; Zhou J; Shoufeng Y; Brindley MA; Zhang B-Z; Huang J-D; To KK-W; Yuen K-Y; Chu H Nature 2022, 603, 693–699. [PubMed: 35062016]

(18). Zhang L; Li Q; Liang Z; Li T; Liu S; Cui Q; Nie J; Wu Q; Qu X; Huang W; Wang Y Emerging microbes & infections 2022, 11, 1–5. [PubMed: 34890524]

(19). Liu L; Iketani S; Guo Y; Chan JF; Wang M; Liu L; Luo Y; Chu H; Huang Y; Nair MS; Yu J; Chik KK-H; Yuen TT; Yoon C; To KK; Chen H; Yin MT; Sobieszczyk ME; Huang Y; Wang HH; Sheng Z; Yuen K-Y; Ho DD Nature 2021, 1–8.

(20). Lu L; Mok BW-Y; Chen L; Chan JM-C; Tsang OT-Y; Lam BH-S; Chuang VW-M; Chu AW-H; Chan W-M; Ip JD; Chan B; Zhang R; Yip C; Cheng V; Chan KH; Jin DY; Hung I; Yuen; Yung K; Chen H; To KKW Clin Infect Dis, doi:10.1093/cid/ciab1041 2021.

(21). Chen J; Wei G-W arXiv:2202.05031 2022, DOI: 10.48550/arXiv.2202.05031.

(22). Chen J; Qiu Y; Wang R; Wei G-W arXiv preprint arXiv:2205.00532 2022.

(23). Chen J; Wang R; Hozumi Y; Liu G; Qiu Y; Wei X; Wei G-W arXiv preprint arXiv:2210.09485 2022.

(24). Li L; Liao H; Meng Y; Li W; Han P; Liu K; Wang Q; Li D; Zhang Y; Wang L; Fan Z; Zhang Y; Wang Q; Zhao X; Sun Y; Huang N; Qi J; Gao GF Cell 2022, 185, 2952–2960. [PubMed: 35809570]

(25). Lan J; Ge J; Yu J; Shan S; Zhou H; Fan S; Zhang Q; Shi X; Wang Q; Zhang L; Wang X Nature 2020, 581, 215–220. [PubMed: 32225176]

(26). Mannar D; Saville JW; Zhu X; Srivastava SS; Berezuk AM; Tuttle KS; Marquez AC; Sekirov I; Subramaniam S Science 2022, 375, 760–764. [PubMed: 35050643]

(27). For Disease Control, C.; Prevention Variant Proportions, https://covid.cdc.gov/covid-data-tracker/#variant-proportions, 2021.

(28). Jankauskaitė J; Jiménez-Garcia B; Dapkūnas J; Fernández-Recio J; Moal IH Bioinformatics 2019, 35, 462–469. [PubMed: 30020414]

(29). Chan KK; Dorosky D; Sharma P; Abbasi SA; Dye JM; Kranz DM; Herbert AS; Procko E Science 2020, 369, 1261–1265. [PubMed: 32753553]

(30). Starr TN; Greaney AJ; Hilton SK; Ellis D; Crawford KH; Dingens AS; Navarro MJ; Bowen JE; Tortorici MA; Walls AC; King NP; Veesler D; Bloom JD Cell 2020, 182, 1295–1310. [PubMed: 32841599]

(31). Linsky TW; Vergara R; Codina N; Nelson JW; Walker MJ; Su W; Barnes CO; Hsiang T-Y; Esser-Nobis K; Yu K; Reneer ZB; Hou YJ; Mason ML; Chen J; Chen A; Berrocal T; Peng H; Clairmont NS; Castellanos J; Lin Y-R; Josephson-Day A; Baric RS; Fuller DH; Walkey CD; Ross TM; Swanson R; Bjorkman PJ; Gale M Jr.; Blancas-Mejia LM; Yen H.-l.; Silva D-A Science 2020, 370, 1208–1214. [PubMed: 33154107]

(32). Starr TN; Greaney AJ; Stewart CM; Walls AC; Hannon WW; Veesler D; Bloom JD bioRxiv 2022.

(33). Zomorodian A; Carlsson G Discrete Comput Geom 2005, 33, 249–274.

(34). Edelsbrunner H; Harer J Contemp. Math 2008, 453, 257–282.

(35). Wang R; Nguyen DD; Wei G-W International journal for numerical methods in biomedical engineering 2020, 36, e3376. [PubMed: 32515170]

(36). Altschul SF; Madden TL; Schäffer AA; Zhang J; Zhang Z; Miller W; Lipman DJ Nucleic acids research 1997, 25, 3389–3402. [PubMed: 9254694]

(37). Wee J; Xia K Briefings in Bioinformatics 2022, 23.

(38). Shu Y; McCauley J Eurosurveillance 2017, 22, 30494. [PubMed: 28382917]

(39). Wu F; Zhao S; Yu B; Chen Y-M; Wang W; Song Z-G; Hu Y; Tao Z-W; Tian J-H; Pei Y-Y; Yuan M-L; Zhang Y-L; Dai F-H; Liu Y; Wang Q-M; Zheng J-J; Xu l.; Holmes EC; Zhang Y-Z Nature 2020, 579, 265–269. [PubMed: 32015508]

(40). Yin C. Genomics 2020, 112, 3588–3596. [PubMed: 32353474]

(41). Kim S; Misra A Annu. Rev. Biomed. Eng 2007, 9, 289–320. [PubMed: 17391067]

(42). Of Medicine, N. N. L. SARS-CoV-2 Variants Overview, https://www.ncbi.nlm.nih.gov/activ, Accessed: 2022-10-16.

(43). covSPECTRUM Detect and analyze variants of SARS-CoV-2, https://cov-spectrum.org/explore/World/AllSamples/Past6M/, Accessed: 2022-10-16.

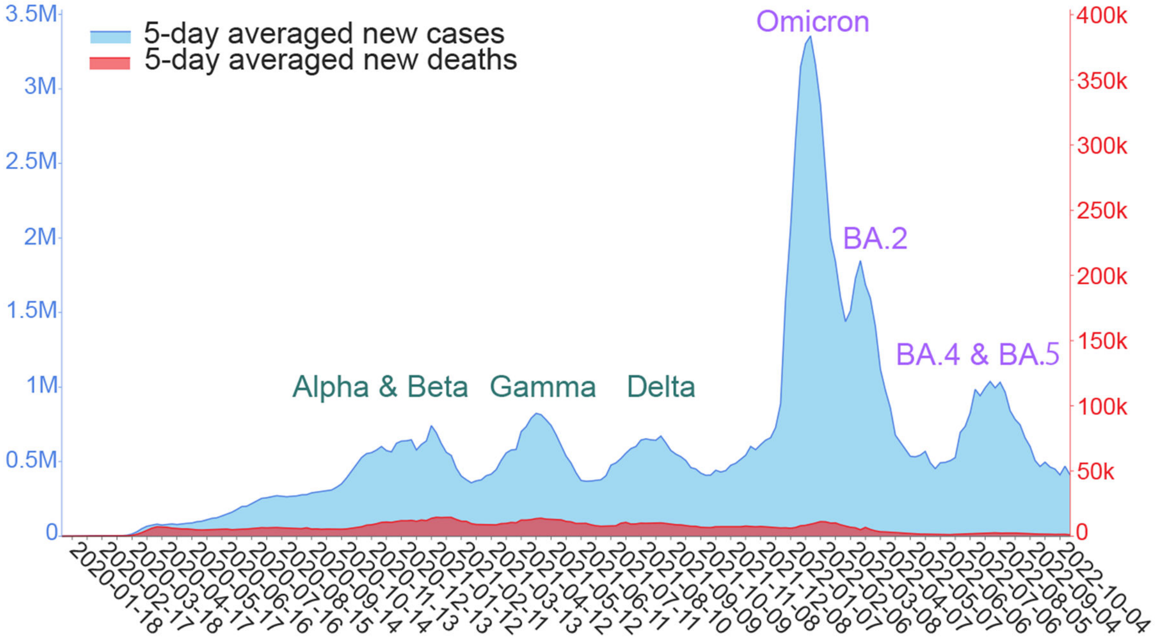


Figure 1:

Illustration of six waves of daily COVID-19 cases (light blue) and deaths (red) driven by dominant SARS-CoV-2 variants since 2020 [1]. The curves are smoothed by five-day averages.

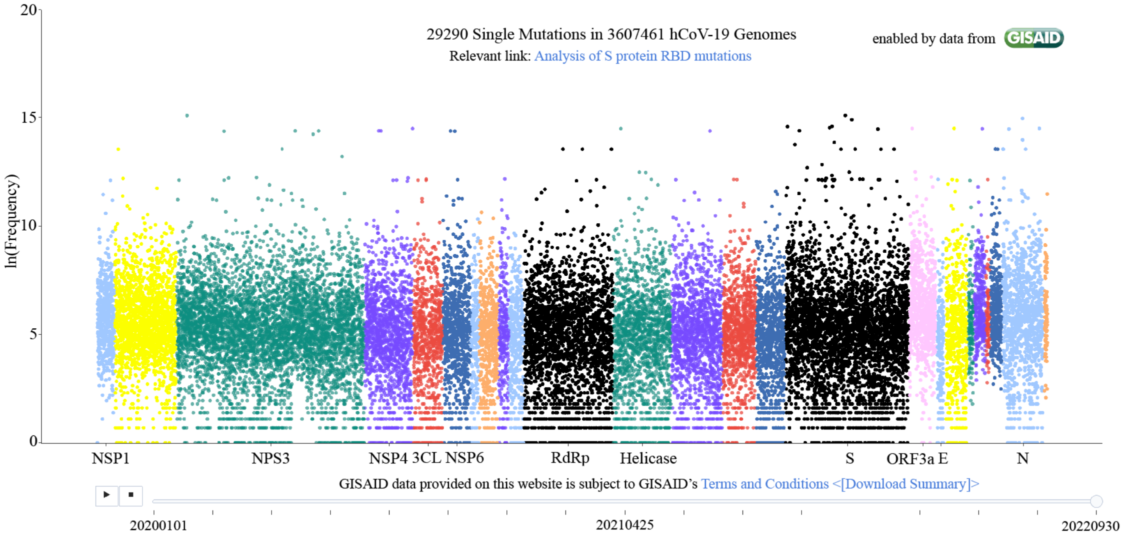


Figure 2:

Illustration of unique mutations on SARS-CoV-2 genomes extracted from patients. Each dot represents a unique mutation. The *x*-axis is the gene position of a mutation and the *y*-axis represents its observed frequency in the natural logarithmic scale.

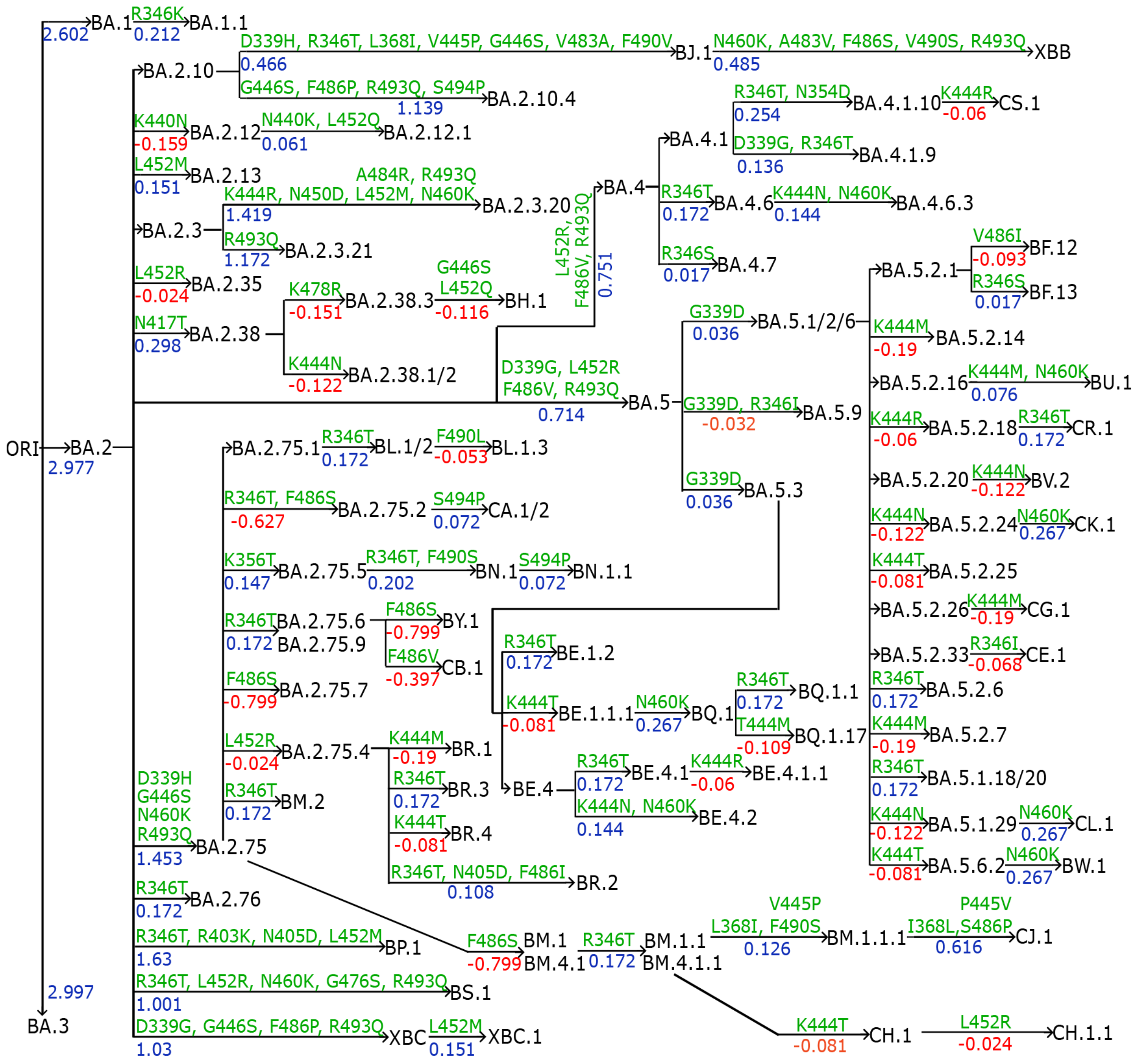


Figure 3:

Annotation tree plot of 106 newly occurred Omicron subvariants. BFE changes (kcal/mol) are marked from parent generations to children as well as mutations.

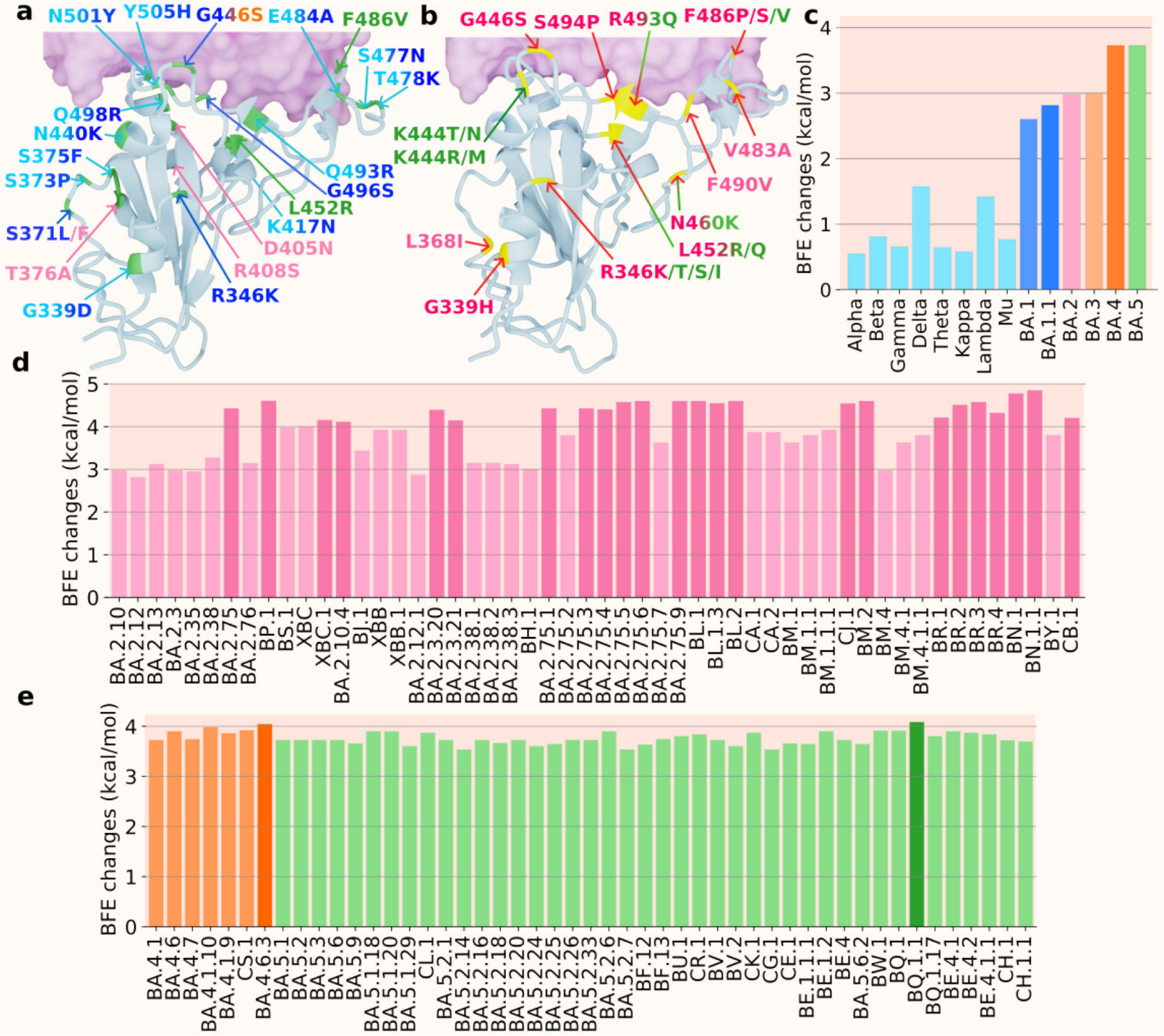


Figure 4:

a. and **b**. the 3D structure of BA.2 (PDB: 7XB0 [24]) with two sets of mutations (colors are consistent with those in c and integrated colors indicate that mutation appears on multiple variants). **a**. the mutations of precious VOCs (in cyan) and BA.1 (in blue), BA.2 (in pink), BA.3 (in orange), BA.4, and BA.5 (in green). **b**. the mutations of the Omicron subvariants with BA.2 sublineages (pink) and BA.5 sublineages (in green). **c**. A comparison of predicted mutation-induced BFE changes for previous VOCs and Omicron subvariants. Previous VOCs (in cyan): Alpha, Beta, Gamma, Delta, Theta, Kappa, Lambda, and Mu; BA.1 and BA.1.1 (in blue); BA.2 (in pink); BA.3 (in light orange); BA.4 (in orange); BA.5 (in green). **d** BA.2 sublineages (in pink) **e** BA.4 sublineages (in orange) and BA.5 sublineages (in green).

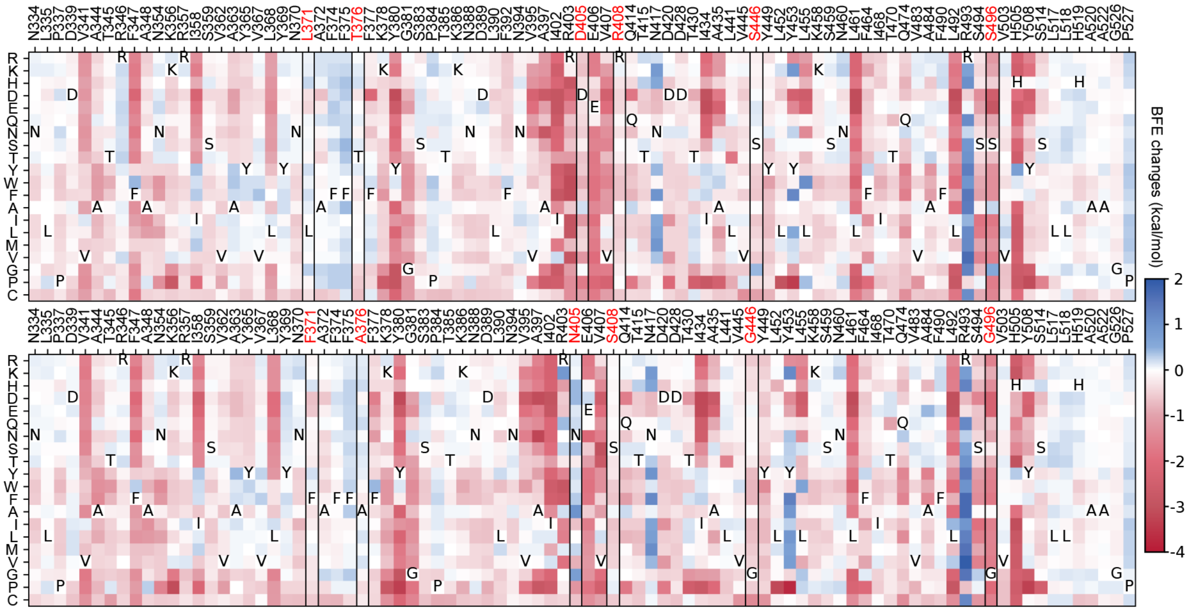


Figure 5:

Heatmap of mutation-induced BFE change predictions of BA.1 (top panel) and BA.2 (bottom panel). Residues that have at least one mutation-induced BFE change greater than 0.1 kcal/mol are selected. The sites of BA.2’s six distinct mutations are marked red and framed in the heatmap.

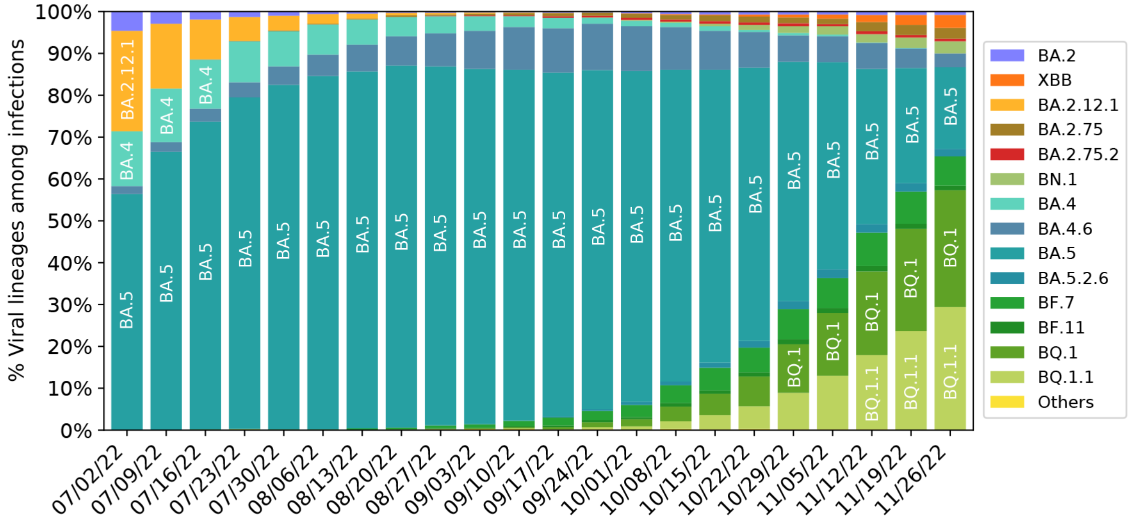


Figure 6:

Weekly viral lineages among infections in the United States from 06/26/2022 to 10/08/2022. AY.1-AY.133, Delta (B.1.617.2), BA.1 and sublineages of BA.1 variant are aggregated to category “Others”. BA.2 sublineages except BA.2.12.1, BA.2.75, BA.2.75.2, BN.1, XBB and their sublineages, are aggregated with BA.2. BA.4 sublineages are aggregated to BA.4 except BA.4.6. Sublineages of BA.5 are aggregated to BA.5 except BF.7, BF.11, BA.5.2.6, BQ.1 and BQ.1.1. The spike substitution R346T is included in lineages BA.2.75.2, XBB, BN.1, BA.4.6, BF.7, BF.11, BA.5.2.6, and BQ.1.1 Data from CDC website [27].