Nucleic Acids Research, 2024, **52**, D701–D713
https://doi.org/10.1093/nar/gkad958
Advance access publication date: 28 October 2023
Database issue

COV2Var, a function annotation database of SARS-CoV-2 genetic variation

Yuzhou Feng[1,2], Jiahao Yi[3], Lin Yang[4], Yanfei Wang[5], Jianguo Wen[5], Weiling Zhao[5], Pora Kim[5,*] and Xiaobo Zhou[5,6,7,*]

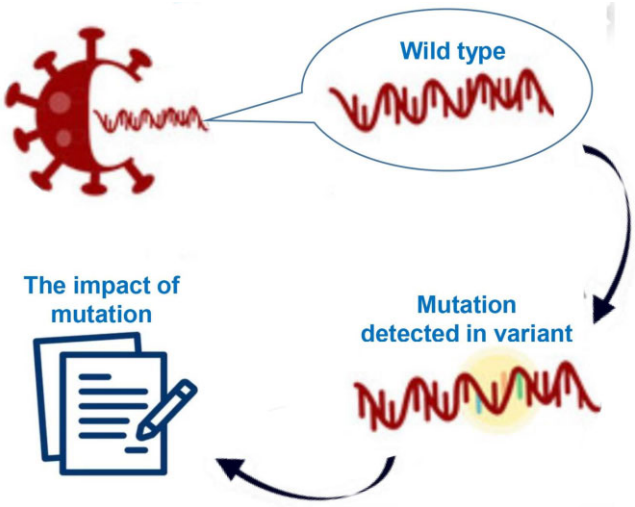
[1]Department of Laboratory Medicine and West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu 610041, China
[2]Med-X Center for Informatics, Sichuan University, Chengdu 610041, China
[3]School of Big Health, Guizhou Medical University, Guiyang 550025, China
[4]Department of Cardiology and Laboratory of Gene Therapy for Heart Diseases, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University and Collaborative Innovation Center for Biotherapy, Chengdu 610041, China
[5]Center for Computational Systems Medicine, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[6]McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[7]School of Dentistry, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[*]To whom correspondence should be addressed. Tel: +1 713 500 3923; Email: Xiaobo.Zhou@uth.tmc.edu
Correspondence may also be addressed to Pora Kim. Tel: +1 713 500 3636; Email: Pora.Kim@uth.tmc.edu

Abstract

The COVID-19 pandemic, caused by the coronavirus SARS-CoV-2, has resulted in the loss of millions of lives and severe global economic consequences. Every time SARS-CoV-2 replicates, the viruses acquire new mutations in their genomes. Mutations in SARS-CoV-2 genomes led to increased transmissibility, severe disease outcomes, evasion of the immune response, changes in clinical manifestations and reducing the efficacy of vaccines or treatments. To date, the multiple resources provide lists of detected mutations without key functional annotations. There is a lack of research examining the relationship between mutations and various factors such as disease severity, pathogenicity, patient age, patient gender, cross-species transmission, viral immune escape, immune response level, viral transmission capability, viral evolution, host adaptability, viral protein structure, viral protein function, viral protein stability and concurrent mutations. Deep understanding the relationship between mutation sites and these factors is crucial for advancing our knowledge of SARS-CoV-2 and for developing effective responses. To fill this gap, we built COV2Var, a function annotation database of SARS-CoV-2 genetic variation, available at http://biomedbdc.wchscu.cn/COV2Var/. COV2Var aims to identify common mutations in SARS-CoV-2 variants and assess their effects, providing a valuable resource for intensive functional annotations of common mutations among SARS-CoV-2 variants.

Graphical abstract



Received: August 15, 2023. **Revised:** September 29, 2023. **Editorial Decision:** October 12, 2023. **Accepted:** October 16, 2023

Introduction

As of July 2023, SARS-CoV-2, the virus responsible for causing COVID-19 pandemic, has resulted in >450 million infections and 6 million deaths worldwide (https://covid19.who.int/). Like other RNA viruses, SARS-CoV-2 has a high mutation rate (1). Although most mutations are functionally neutral and occur randomly, some mutations can confer a survival advantage to the virus. Viral mutations can affect host-pathogen interactions in many ways (i.e. affect viral spread, impact virulence, escape natural or vaccine-induced immunity, evade therapies or detection by diagnostic tests and change host species range) (2,3). For example, the D614G (S) mutation in the spike protein has been linked to increased transmission efficiency (4,5). The virus employs a survival strategy by gradually accepting new mutations at a slow pace, opting instead to combine existing mutations for evolutionary advantage (6). This suggests that the virus may have already explored the majority of advantageous mutations, leading it to continue its evolution by utilizing a combination of existing mutations (6). Therefore, it is essential to investigate the exact impact of individual mutations, aiming to ascertain whether the new mutation possesses characteristics that could make it more lethal and contagious.

To date, there are multiple resources related with SARS-CoV-2 mutation annotation (i.e. COVID-19 CG (7), CovMT (8), CoV-GLUE (9), coronapp (10), GESS (11) and Outbreak.info (12)). However, these resources provide a list of mutations detected in SARS-CoV-2 variants without intensive functional annotations and new insights into human health (Supplementary Table S1). To address this gap, we performed comprehensive bioinformatics analyses for over 13 billion SARS-CoV-2 genome sequences and associated metadata. Through this extensive examination, we have identified a total of 9832 common mutations in the SARS-CoV-2 genome. Then, we conducted intensive annotation of the functional mechanisms underlying 9832 common mutations. For example, we examined the relationship between these mutations and various factors such as disease severity, patient age, patient gender, cross-species transmission, viral immune escape, immune response level, viral transmission capability, viral evolution, viral protein structure, viral protein function, viral protein stability and concurrent mutations. The resulting knowledgebase, COV2Var, available at http://biomedbdc.wchscu.cn/COV2Var/, provides an important resource for routine SARS-CoV-2-related research. COV2Var will be helpful to enhance our understanding of SARS-CoV-2, its adaptive mechanisms in new environments and the development of effective strategies to combat the virus.

Data integration and annotations

Data collection and quality control

On 2 March 2023, we retrieved SARS-CoV-2 sequences from the Global Initiative on Sharing Avian Influenza Data (GISAID) database (13), covering sequences collected between December 2019 and February 2023. We applied filters to select sequences with complete genomes, excluding those with low coverage (>5% undefined nucleotide 'N') and ensuring complete collection date information. Additionally, we collected metadata associated with each sequence. We categorized the sequences by variant (such as Alpha, Delta) using the Pango nomenclature SARS-CoV-2 lineage designation of each sequence (14). This categorization excluded records designated as 'None' or 'Unassigned.'

Mutational analysis

For mutation analysis of the SARS-CoV-2 genome sequence, we employed the microbial genomics mutation tracker (MicroGMT) software with default annotations for SARS-CoV-2 (Wuhan-Hu-1) (15). MicroGMT takes assembled genome sequences as input and compares them to a reference sequence to detect and characterize insertions, deletions and point mutations. The reference sequence used for analysis was the SARS-CoV-2 isolate Wuhan-Hu-1 (GenBank accession NC_045512.2) (16). All nucleotide position labels in our analysis were based on alignment with this reference sequence. In our study, we utilized this tool to analyze the genome sequences of an extensive 13 344 494 SARS-CoV-2 isolates. We applied filtering to identify more prevalent and biologically significant mutations with potential implications for viral adaptation and survival advantage. The filtering criteria were as follows: (i) a mutation has a frequency >0.01 in at least one of the 2735 viral lineages (Pango lineage) and occurs at least twice within that specific lineage; (ii) a mutation has to be present in two or more lineages out of a total of 2735 lineages.

Geographical and temporal distributions analysis of mutations and frequency across lineages

The metadata for each SARS-CoV-2 genome sequence includes location, sampling time and Pango lineage information. By combining the metadata and detected mutations from the sequences, we gained valuable insights into the mutation patterns, including their geographical and temporal distributions and their frequency across different lineages. The distribution and alternation of mutations worldwide were visualized using Python's pyecharts library. Through these efforts, we were able to unveil the mutation landscape and visually track their changes over time for each mutation.

Investigating the relationship between mutations and alternative non-human animal hosts

The metadata for each SARS-CoV-2 genome sequence includes host information. We analyzed the distribution of mutations in different non-human animal hosts based on the sequence host information to undergo cross-species transmission. We retained the mutation that appears in at least three non-human animal hosts' sequences.

Investigating the association between mutations and age, gender and patient status

We employed a logistic regression model to investigate the association between mutations and patient age, patient gender and patient status. The logistic regression model was conducted using the glm function in R (17). In detail, we categorized the data into different age groups (0–17, 18–39, 40–64, 65–84 and 85+), genders (female and male) and patient statuses (ambulatory, deceased, homebound, hospitalized, mild and recovered) based on GISAID classifications. In the analysis of the association between mutations and patient status, the model included mutation, patient status, patient age, gender and the origin and collection time of gene sequences. Similarly, in the analysis of

the association between mutations and patient demographic information (age and gender), the model incorporated mutation, patient age, gender and the origin, and collection time of gene sequences. A *P*-value <0.001 for the mutation variable indicates that the observed differences in the fraction of pre-mutation and after-mutation sequences are statistically significant, providing evidence against the null hypothesis of no difference.

Analyzing the link between mutation sites and natural selection

Genes containing adaptive mutations in the genome are constantly increasing due to the predominance of positive natural selection. We used the RASCL pipeline with default protein annotations for SARS-CoV-2 (Wuhan-Hu-1) to investigate the natural selection pressures acting on SARS-CoV-2 genes in each lineage sequence (18). In detail, we conducted analyses using advanced molecular evolution models from the HyPhy (19), including fixed effects likelihood (FEL) (20), mixed effects model of evolution (MEME) (21) and fast unbiased Bayesian approximation (FUBAR) (22) to detect sites undergoing positive selection or negative selection. In the case of FEL and MEME, sites with *P*-value <0.05 and in FUBAR, posterior probability >0.95 is considered significantly positive.

Analyzing the impact of the mutations on protein physicochemical properties

The reference protein sequences were obtained from GenBank (GenBank accession NC_045512.2) (16). Point mutation protein sequences were generated by replacing the corresponding amino acids at specific positions. Subsequently, the physicochemical properties of the protein sequence were analyzed using the Expasy ProtParam server (https://web.expasy.org/protparam/) (23). Various physicochemical properties were examined, including the number of amino acids, molecular weight, theoretical isoelectric point (p*I*), extinction coefficients, aliphatic index and grand average of hydropathicity (GRAVY). We compared the changes in physicochemical properties before and after the mutation and considered changes exceeding 10% as significant alterations.

Analyzing the impacts of mutations on protein stability and function

The I-Mutant 2.0 web server (https://folding.biofold.org/i-mutant/i-mutant2.0.html) was used to predict changes in the stability of mutated protein sequences (24). The *in vitro* environment was simulated using pH of 7 and temperature 25°C, while the *in vivo* environment was simulated using pH 7.4 and temperature 37°C. The resulting ΔDDG values represent the predicted energy change. Based on this prediction, I-Mutant determines whether a specific mutation will increase (ΔDDG > 0) or decrease (ΔDDG < 0) the stability of the protein. MutPred2 was used with default parameters to offer probabilistic insights into the pathogenicity of amino acid substitutions (25). A score >0.5 indicates an increased likelihood of pathogenicity.

Analysis of intrinsically disordered regions (IDRs)

The reference protein sequences of SARS-CoV-2 were obtained from GenBank (GenBank accession NC_045512.2) (16). The fraction of disordered residues was calculated using long disorder prediction mode of the IUPred3 with default (medium smoothing) parameters (26). Residues with a predicted score >0.5 are considered disordered.

Alterations in enzyme cleavage sites induced by mutations

The reference protein sequences were obtained from GenBank (GenBank accession NC_045512.2) (16). Point mutation protein sequences were generated by replacing the corresponding amino acids at specific positions. Subsequently, the protein sequences were analyzed for potential cleavage sites using the PeptideCutter tool on the ExPASy Server (https://web.expasy.org/peptide_cutter/), considering all available enzymes and chemicals (23). We compared the changes in the enzyme cleavage sites before and after the mutation and retained the sites that have undergone changes.

Prediction of antigenicity and immunogenicity of spike protein mutations

The spike protein sequences were obtained from GenBank (GenBank accession NC_045512.2) (16). Point mutation protein sequences were generated by replacing the corresponding amino acids at specific positions. The antigenicity of spike protein was predicted using the VaxiJen 2.0 server (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html) with 'virus' selected as the target organism (27). Antigens with a prediction score of more than 0.5 are considered candidate antigens. The MHC-I immunogenicity of the spike protein mutation was checked by IEDB Class I immunogenicity tool (http://tools.iedb.org/immunogenicity/) with default (1st, 2nd and C-terminus amino acids) parameters (28). MHC I immunogenicity score >0 indicates a higher likelihood of stimulating an immune response. For a mutation that significantly influences antigenicity or immunogenicity, we considered mutations with an absolute change in antigenicity or immunogenicity score more than three times the median absolute change across all sites.

Analyzing the impacts of mutations on viral transmissibility by altering the affinity between receptor binding domain (RBD) and ACE2 receptor

We used the results of the deep mutational scanning (DMS) approach (29–31) to experimentally measure how all possible SARS-CoV-2 RBD amino acid mutations affect ACE2-binding affinity. A positive Δbinding affinity value ($\Delta\log_{10}$(KD, app) > 0) signifies an increased affinity between RBD and ACE2 receptor due to a mutation. Conversely, a negative value ($\Delta\log_{10}$(KD, app) < 0) indicates a reduced affinity between RBD and ACE2 receptor caused by mutations. The *t*-test was used to determine the significance of this change, with a *P*-value < 0.05 indicating a significant alteration in the affinity between RBD and ACE2 after the mutation.

Analyzing the impacts of mutations on immune escape by altering the affinity between RBD and antibody/serum

Deep mutational scanning can systematically measure the antigenic impacts of all possible amino-acid mutations in the critical regions of the spike on monoclonal antibodies or sera. We utilized the results of deep mutational scanning experi-

ments to prospectively measure how viral mutations impact antibody binding or neutralization (32–34). The data we collected includes RBD antibody Classes 1–4 (Class 1—ACE2 blocking with binding to open RBD only, Class 2—ACE2 blocking with binding to both open and closed RBD, Class 3—non-ACE2 blocking with binding to both open and closed RBD and Class 4—non-ACE2 blocking with binding to open RBD only), convalescent plasma and post-vaccination sera to examine the impact of mutation sites on immune evasion. A value of 0 indicates the variant always binds to the antibody, and a value of 1 means that the variant always escapes antibody binding. A mutation with an escape score greater than 0.1 (10% of the maximum score 1) is classified as an escape.

Analyzing the co-mutation patterns of SARS-CoV-2

We calculated the correlation between two mutation sites across 2735 viral lineage categories, aiming to explore their degree of association. Each sequence contains different mutation site information. First, we processed the sequences and constructed a pivot table. In the pivot table, each row represents a sample, and each column represents a mutation site. The values in the table are either 0 or 1, indicating whether each sample has the corresponding mutation. Finally, we calculated the Spearman correlation coefficients between different mutation sites. Holm–Bonferroni method was used for multiple test adjustments (35). We retained mutation pairs with correlation values >0.6 or <-0.6 and Holm–Bonferroni corrected P-values <0.05.

Manual curation of mutation-related literature from PubMed

On 29 July 2023, we searched PubMed to retrieve the related literature related to the mutation. Taking D614G (S) as an example, the following query was used: (‘COVID-19’[Title/Abstract] OR ‘SARS-CoV-2’[Title/Abstract]) AND (‘1841A > G’[Title/Abstract] OR ‘Asp614Gly’[Title/Abstract] OR ‘D614G’[Title/Abstract]). Results were manually filtered for relevance.

Web interface and analysis results

Database overview

In this study, we compiled a vast dataset of over 13 billion SARS-CoV-2 genome sequences from GISAID, along with relevant metadata for each sequence (13). This extensive collection of genomes spans the timeframe from December 2019 to February 2023. These genome sequences represent a rich diversity of 2735 viral lineages from 218 distinct geographical regions originating across 35 different host species, as depicted in Figure 1A and B. The metadata file encompasses crucial information, including region of origin, date of collection, date of submission and lineage, as well as the status, age and gender of the host.

We analyzed mutations in each SARS-CoV-2 genome and identified 9832 common mutations after filtering. Compared with other mutations, these mutations exhibited pronounced adaption and transmission, indicating their advantageous role in evolutionary and biological significance. The distribution of these mutations across the SARS-CoV-2 genome is illustrated in Figure 1C, and the detailed list of mutations can be found in Supplementary Table S2. We then performed extensive functional annotation by integrating sequence information and metadata for 9832 common mutations.

The overall schema of COV2Var is represented in Figure 2. For 9832 individual SARS-CoV-2 mutations, we investigated their distribution across 218 geographical regions, temporal variations and frequency patterns within 2735 lineages to discern their impact on evolutionary dynamics. We also investigated the association between mutations and patients of different ages, genders and statuses. Among the 9832 mutations analyzed, 429 were related to patient gender, 1593 to patient status and 2762 to patient age. By analyzing the relationship between mutations and alternative non-human animal hosts (35 different species), we revealed the impact of mutations on viral adaptation and cross-species transmission. A total of 1658 mutations were detected in non-human animal hosts. Investigating the impact of mutations on protein stability, function mechanism, enzyme cleavage sites and physicochemical properties provides valuable insights into their structural and functional implications. Exploring the impact of spike protein mutations on antigenicity, and immunogenicity provides insights into the immune response. Unraveling the impact of mutations on the interaction between the RBD and ACE2 receptor using DMS experimental data provides insights into their effects on viral transmissibility. Among them, 44 mutations have been identified to increase the binding affinity between the RBD and the ACE2 receptor. Unraveling the impact of mutations on the interaction between the RBD and antibodies/serum using DMS experimental data to understand evasion of host immune response. A total of 1504 antibodies/serum data were collected for this analysis. We investigated the co-mutation patterns of SARS-CoV-2 across 2735 viral lineages to unravel the cooperative effects of different mutations on genetic variation and disease phenotypes. We curated 9832 common mutations through the PubMed search. Among these, 615 mutations have been reported in 2587 papers, associating with a critical role in COVID-19 progression.

Association of mutations with age, gender and patient status

Previous studies have reported associations between specific mutations and changes in patient age, gender distribution and patient status. In this study, we integrated mutations and patient information, encompassing age, gender and patient status, to better understand these potential correlations. Among the 9832 mutations analyzed, we identified 429 mutations associated with patient gender, 1593 mutations associated with patient status and 2762 mutations associated with patient age. For example, mutations like P4715L (ORF1ab) and D614G (S) have been linked to higher fatality rates (36–38). Conversely, P13L (N), Y789Y (S), L37F (ORF10) and L6420L (ORF1ab) are speculated to correlate with reduced mortality rates (39,40). Additionally, S24L (ORF8) appears more frequent among females (41). The D614G (S) mutation was found to be associated with male gender (42,43) and advanced age (43,44). These observed trends are consistent with our results.

Natural selection sites

The detection and quantification of evolutionary pressure revealed that the majority of SARS-CoV-2 codons were under strong positive or negative selection. Positive selection can gradually increase specific mutations within the viral pop-

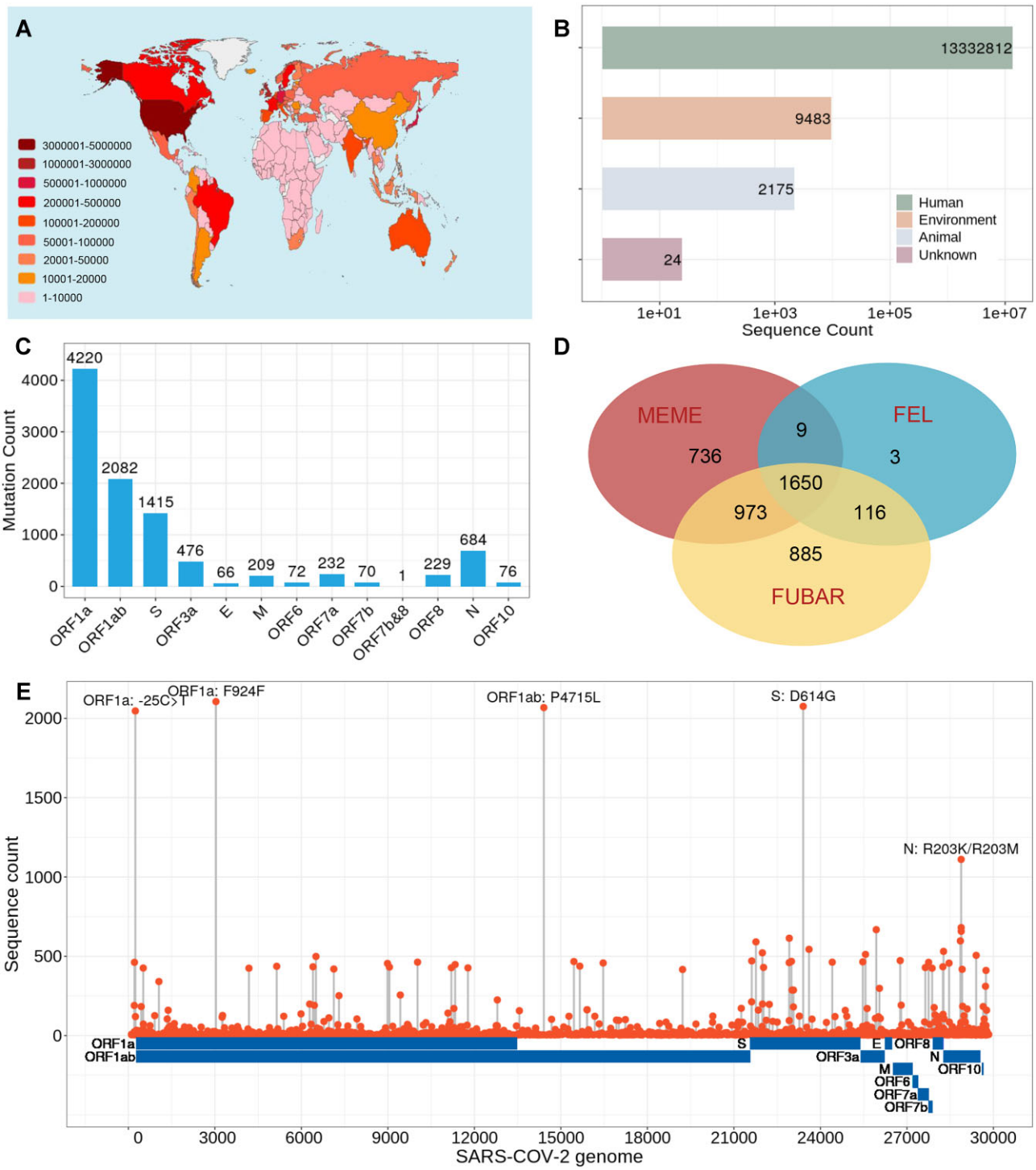


Figure 1. Overview of data distribution. (**A**) Global distribution of sequences across 218 regions. (**B**) Source origins of all genome sequences. (**C**) Distribution of 9832 mutations in the SARS-CoV-2 genome. The y-axis represents the number of distinct mutations. (**D**) Positively selected mutations were detected by FEL, MEME and FUBAR methods among 9832 mutations. (**E**) Distribution of mutations from non-human animal hosts across the SARS-CoV-2 genome. The y-axis represents the number of sequences carrying each mutation.

ulation, as these mutations may enhance the virus's adaptability, transmissibility and drug resistance. Here, we conducted selection pressure analysis on 2735 lineages individually and identified a total of 4372 sites subjected to positive selection pressure. Among these, MEME identified 3368 sites, FEL identified 1778 sites, FUBAR identified 3624 sites and a combination of FEL, MEME and FUBAR identified 1650 sites (Figure 1D). Mutation under positive selection favor the virus's survival. For example, the positions L18 (S), L382 (ORF1ab), K417 (S), N501 (S), H655 (S) and P681 (S) have previously been reported to be under positive selection and may have potentially significant fitness impacts (45,46). These adaptive mutations are generally potential drug targets.

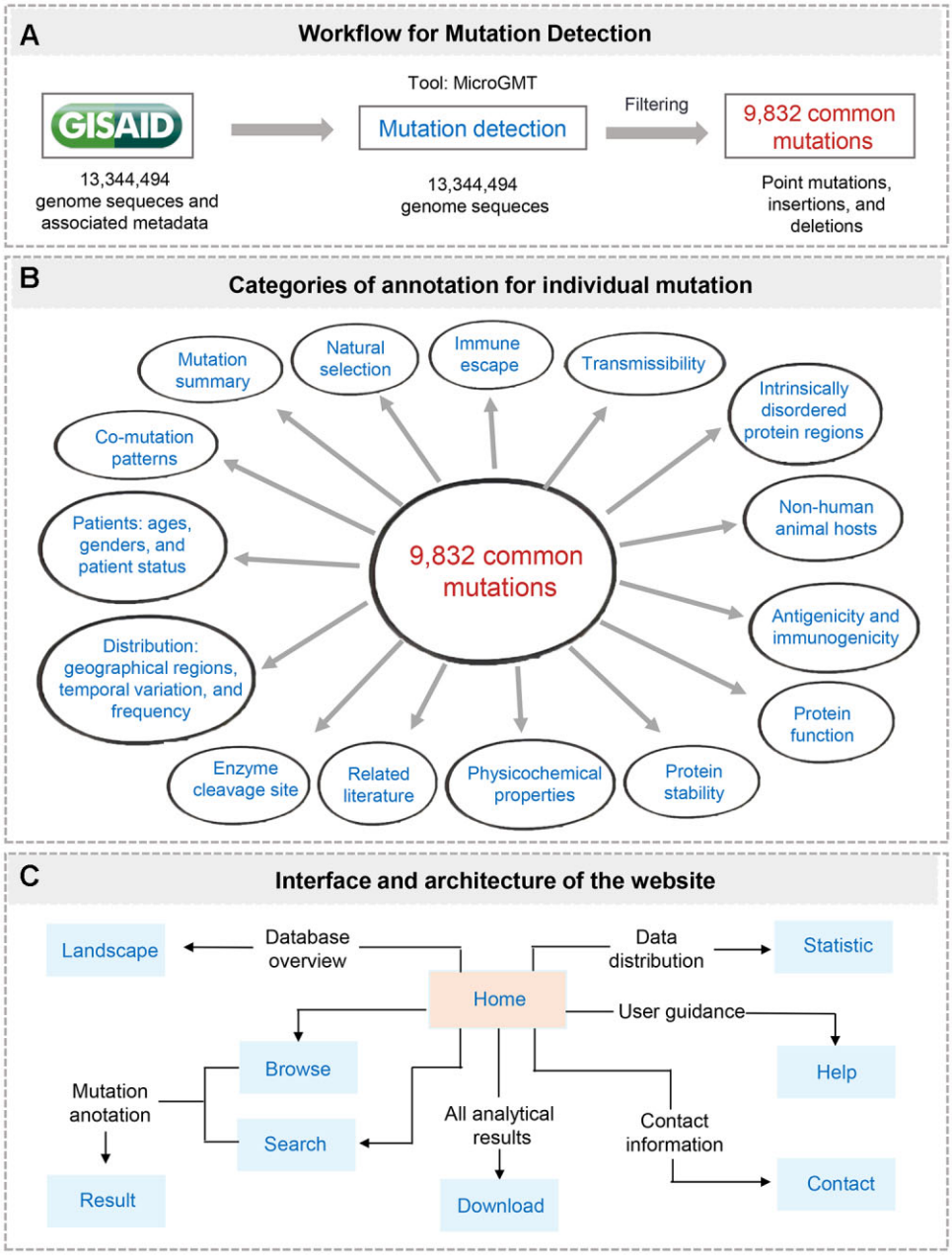


Figure 2. COV2Var pipeline overview. (**A**) Mutation analysis workflow leading to 9832 common mutations. (**B**) Main categories of common mutations among 9832 Individuals. (**C**) Interface and architecture of the COV2Var website.

Distribution in non-human animal hosts

Current, research indicates that SARS-CoV-2 can undergo cross-species transmission. Several cases of animal contamination by different human variants of SARS-CoV-2 have been reported, affecting a wide range of species, including pets, livestock and wildlife (47,48). Contamination of hamsters has been reported resulting in subsequent transmission back to humans, potentially triggering an epidemic that spreads via human-to-human transmission (49). In this study, we investigate the distribution of mutations in non-human animal hosts to gain insights into the possibility of cross-species transmission events. A total of 1658 mutation sites were identified in non-human animal hosts, spanning 29 distinct species. The genomic distribution of these 1658 mutations is shown in Figure 1E. Among them, several mutations, such as Y453F (S), F486L (S) and N501T (S), have been reported to facilitate the adaptation of SARS-CoV-2 to infect mustelids (50,51). Cross-species virus transmission plays a pivotal role in the evolution and emergence of variants. Analyzing mutations in non-human animals offers insights into the potential for cross-species transmission and enhances early warning capabilities.

Distribution of mutation on IDRs

Intrinsically disordered regions refer to protein regions with no unique 3D structure. These proteins' highly dynamic disordered regions have been linked to important phenomena, including enzyme catalysis and allosteric regulation, as well as crucial physiological functions like cell signaling and transcription (52). In viral proteins, mutations in the disordered regions are critical for immune evasion and antibody escape,

suggesting potential additional implications for vaccines and monoclonal therapeutic strategies (52,53). In total, 9533 mutation sites were found within protein-coding regions (across 12 GenBank protein sequences) (16). Among these, 735 sites are situated within intrinsically disordered regions, distributed across five protein s: ORF1a (15.8%), ORF1ab (0.8%), S (10.6%), ORF3a (4.1%) and N (68.7%). This is consistent with the abundance of IDRs in the nucleoprotein of SARS-CoV-2 (54). IDRs play a special role in increasing binding affinity and enhancing the binding allosteric, enabling the N protein to bind RNA with high cooperativity (55). Azad's study shows that there are a large number of mutations in IDRs of N-protein and multiple mutations lead to considerable alterations in the function of this protein (56). Gaining insights into the structure of Intrinsically disordered regions would provide valuable knowledge for high-throughput screening to identify significant mutation sites that are relevant to biological processes and functions.

Associations of mutations with protein stability and function

A total of 5348 missense mutations were subjected to protein stability and function analysis. The I-Mutant was used to predict the potential impact of a mutation on protein stability. The protein stability analysis indicated that 1029 mutations enhanced protein stability in both *in vitro* and *in vivo* environments. Conversely, 4175 mutations reduced protein stability in both *in vitro* and *in vivo* environments. For example, mutations like L452R (S) and N501Y (S) contribute to the increased stability of the S protein (57,58). Our results are consistent with these studies. Functional analysis was conducted for all the point mutations using MutPred. A total of 254 mutations were found to impact protein functionality.

Associations of mutations with antigenicity and immunogenicity

Besides the critical role of spike protein in viral entry, it can also stimulate the immune response during the viral infection (59). Research has revealed the potential of the spike glycoprotein as an antigenic region (60). Out of the 671 S protein mutations, 99 exhibited significantly altered antigenicity and 27 displayed notable changes in immunogenicity. Specifically, a total of 99 mutations had an antigenicity score change of >0.0102 (3 times the median across sites). Twenty-seven mutations had an immunogenicity score change of >0.2754 (3 times the median across sites).

Impact of mutations on the affinity between RBD and ACE2 receptor

The spike protein mediates the entry of the virus into host cells by binding to ACE2 via a receptor-binding domain (RBD). A total of 39 mutation sites were identified to exhibit alterations in the binding affinity between RBD and ACE2. Mutations in RBD amino acids K417 (S), E484 (S), L452 (S), F486 (S), Y489 (S), Q493 (S), N501 (S) and Y505 (S) residues were found to enhance the affinity of this protein with ACE2 receptor (61–64). Our results are consistent with these studies. Deep mutational scanning holds predictive value for assessing changes in ACE2-binding caused by mutations, facilitating a better understanding of viral transmission capability.

Impact of mutations on the affinity between receptor-binding domain of the spike protein and antibody/serum

Approximately 90% of the plasma or serum neutralizing antibody activity has been reported to target the spike receptor-binding domain (RBD) (65). Unfortunately, the rapid evolution of the spike protein has resulted in diminished serum neutralization potency and facilitated evasion from a majority of monoclonal antibodies (66–68). Deep mutational scanning systematically evaluates the impact of amino acid mutations within critical regions of the RBD on monoclonal antibodies or serum. A total of 1504 antibodies/serum data were collected for this analysis and 114 mutations were identified. Among them, 64 mutation sites exhibited robust immune escape potential and immune evasion was observed in at least 10 different antibodies or sera. For example, among them, E484K (S), A475V (S), L452R (S), V483A (S) and F490L (S) have been identified as an escape mutation that emerges during exposure to mAbs or convalescent plasma (69–73). Understanding these immune escape mutations is crucial for assessing the effectiveness of treatments and vaccines against the virus.

Co-mutation patterns

The correlation coefficients can measure the similarity of the variation trends of two mutation sites in the samples, thus indicating whether they co-occur or co-disappear. Correlation analysis of mutation sites helps us understand whether there is a close relationship or interaction between certain mutations. In this study, we computed the correlations between 9832 mutation sites for each of the 2735 lineages. According to a correlation analysis, 9832 mutations correlated with at least one other mutation. In total, we identified 1 193 386 positive correlation pairs and 37 068 negative correlation pairs. Among them, many correlations have been previously reported. For example, R203K (N) and G204R (N) mutations were frequently observed in combination (74), leading to destabilization and reduced overall structural flexibility of the N protein. D614G (S) and P4715 (ORF1ab) exhibited co-occurrence, showing a high correlation between mutations in two distinct proteins (37,75). Synonymous mutation F924F (ORF1a) was observed co-occurring with other mutations, including 241C > T (-25C > T in ORF1a), P4715L (ORF1ab) and D614G (S) (76). Analyzing co-mutation patterns helps understand the relationships of mutual dependence and exclusion among mutations.

Exploring the differences of S protein N501Y and N501T mutations using COV2Var

To date, the WHO has designated five variants of concern (VOCs): Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2) and Omicron (B.1.1.529) (77,78). The N501Y mutation, located in the S protein's RBD, is prevalent across all five VOCs. In our study, we identified two common mutations at the N501 site: N501Y and N501T. Approximately 45.7% of SARS-CoV-2 variants carry the N501Y mutation, while the N501T mutation is found in only about 0.00043% of variants (Figure 3A, B). This indicates the N501Y mutation is relatively common and widespread. We conducted an in-depth investigation to analyze the factors contributing to the varying occurrences of N501Y and N501T.

Notably, within the B.1.604 lineage, we observed concurrent N501T and N501Y mutations, exhibiting a strong

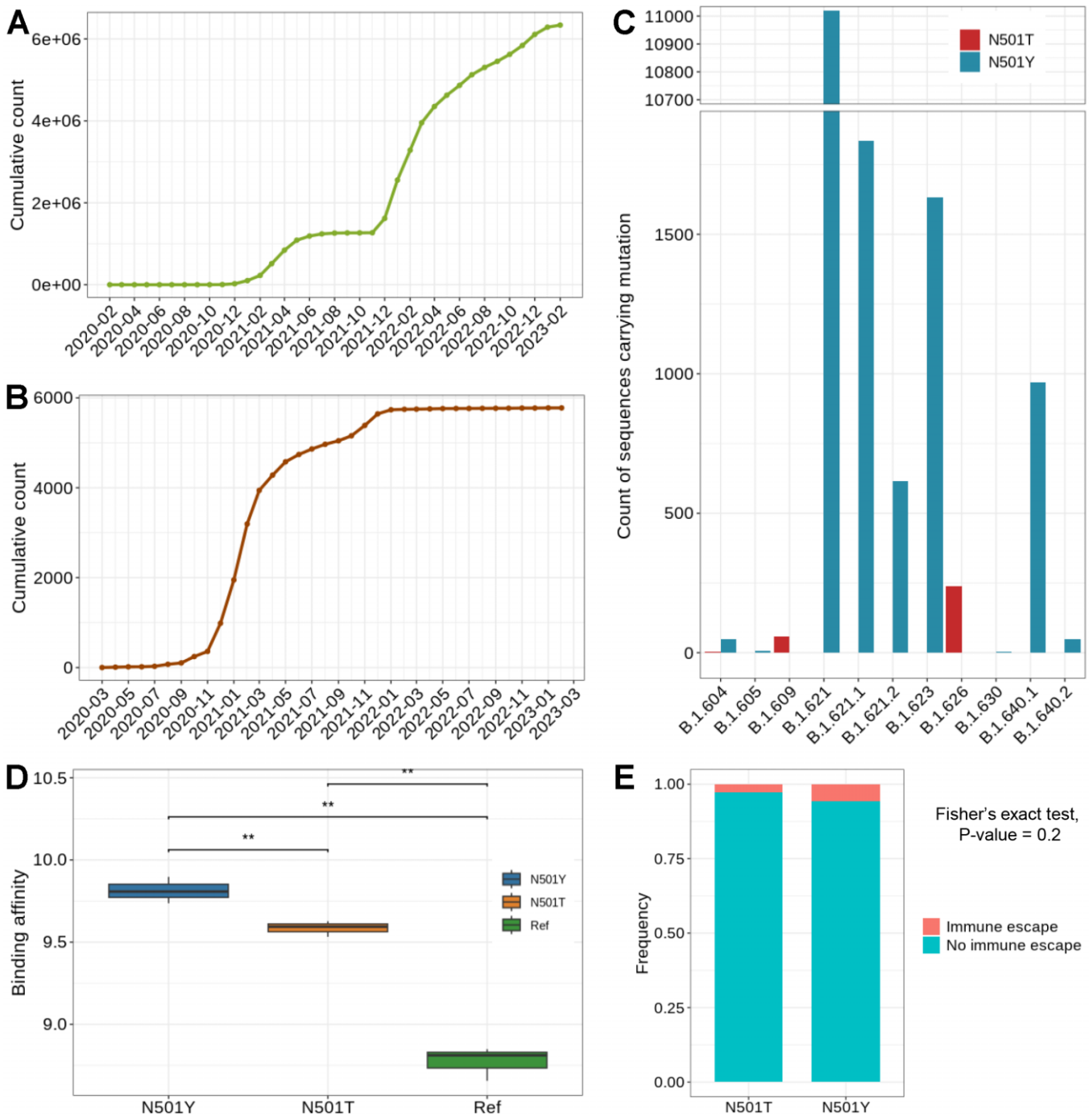


Figure 3. Comparative analysis of N501Y and N501T mutations on SARS-CoV-2 spike protein. (**A**) Temporal dynamics of sequences with the N501Y mutation. (**B**) Temporal dynamics of sequences with the N501T mutation. (**C**) Count of sequences carrying the N501Y or N501T mutation within B.1.6 sub-lineages. (**D**) Comparative ACE2 binding affinity of N501Y and N501T mutations. (**E**) Comparative immune escape of N501Y and N501T mutations.

negative correlation (correlation coefficient: −0.63, Holm–Bonferroni corrected *P*-value < 0.05) between them. N501Y accounts for 87% of the B.1.604 lineage population, while N501T constitutes 5.56% (Figure 3C). We investigated the B.1.6 sub-lineages surrounding the B.1.604 lineage to explore competitive relationships. Sub-lineages within B.1.6 display distinct mutations, some of which carry only N501Y (e.g. B.1.621.1, B.1.605 and B.1.621), others either only carry N501T (B.1.609 and B.1.626), or have both mutations (B.1.604) (Figure 3C). This diverse distribution of N501Y and N501T mutations in B.1.6 sub-lineages was evident. The B.1.621 lineage exhibited the highest prevalence of N501Y. Its frequency ranged from 6.16 to 2212.83 times higher than other sub-lineages carrying N501Y or N501T mutations. In the B.1.621 branch, only one mutation, E484K, showed a strong positive association (correlation coefficient: 0.81, Holm–Bonferroni corrected *P*-value < 0.05) with N501Y, but not N501T. E484K is recognized as an escape mutation emerging during exposure to monoclonal antibodies (mAbs), mAbs combinations and convalescent plasma (69–72). Moreover,

the co-occurrence of N501Y and E484K is frequently observed in VOCs. This suggests that E484K's presence enhances the competitive advantage of N501Y over N501T.

We also explored other factors that could potentially contribute to the differences between the N501Y and N501T mutations. N501Y mutation exhibited enhanced RBD affinity for ACE2 receptor using DMS experimental data (Figure 3D), contributing to increased N501Y transmissibility. Compared to the reference sequence, N501Y maintains unchanged antigenicity and immunogenicity, while N501T displays heightened antigenicity without affecting immunogenicity. There is also no difference in antibody escape capability between N501Y and N501T, as indicated by the Fisher test with a *P*-value >0.05 (Figure 3E). We further integrated sequence metadata for comprehensive analysis. Patients with the N501Y mutation showed significant recovery improvement, from 26.56% to 32.88%. Conversely, patients with N501T mutation experienced poorer outcomes, with mortality rising from 7.14% to 23.08%. N501Y is more prevalent among younger individuals, consistent with prior studies suggesting higher occurrence in youth (79,80). Notably, there is a significant ΔDDG difference between N501Y and N501T, with N501Y being more stable both *in vivo* and *in vitro*. Both N501Y and N501T have been identified in various non-human hosts, aligning with literature and hinting at potential effects on viral adaptation and cross-species transmission. In non-human hosts, sequences carrying the N501Y mutation account for 13.3%, while sequences with the N501T mutation make up 8.4%. This suggests that N501T is more conducive to non-human hosts. The N501T mutation has been reported to promote the adaptation of SARS-CoV-2 for infecting non-human hosts (51). By utilizing COV2Var, we conducted a comparative analysis of the N501Y and N501T mutations, revealing potential reasons for the notable differences in mutation frequency between these mutations in human.

Exploring the latest XBB.1.16* variants with COV2Var

Currently, some lineages and their descendant lineages of the SARS-CoV-2 Omicron XBB variant (e.g. XBB.1.5*, XBB.1.9* and XBB.1.16*) have become predominant worldwide. The '*' symbol represents sub-lineages. For example, XBB.1.5 signifies both XBB.1.5 itself and its associated sub-lineages. The World Health Organization (WHO) has designated XBB.1.5*, XBB.1.9* and XBB.1.16* as variants under monitoring (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants). According to the latest CDC data (https://covid.cdc.gov/covid-data-tracker/#variant-proportions), as of the end of July 2023, XBB.1.16 has emerged as the dominant strain in the USA, rapidly reaching 32.9% prevalence (Figure 4A). This surpasses the proportions of XBB.1.5 (17.2%) and XBB.1.9 (17.6%), indicating a higher potential rate of transmissibility for the XBB.1.16 variant compared to the other two variants. Importantly, the XBB.1.16* variant, as indicated by GISAID's mutation site data (https://gisaid.org/lineage-comparison/) (12,81), exhibits notable characteristic mutations, including L3829F (ORF1a), G18703T (D1746Y in ORF1b or D6147Y in ORF1ab), E180V (S) and T478R (S) (Figure 4B), distinguishing it from the XBB.1.5* and XBB.1.9* variants.

The L3829F mutation is located in the ORF1a gene of the virus. Several studies have reported this mutation, mainly only listing the L3829F mutation without sufficient details on its effects (46,82–84). One study noted that the L3829F is under positive selection, which benefits the virus's survival (46). Our results also support this by showing that the L3829 position is under positive selection. In addition, we found that variants carrying the L3829F mutation are associated with a higher rate of hospitalization upon infection (*P*-value < 0.001). This is consistent with observations on the XBB.1.16*, which has a significantly higher hospitalization rate than other Omicron variants (85). The L3829F mutation enhances the stability of the ORF1a protein both *in vivo* and *in vitro* (ΔDDG > 0). ORF1a gene in SARS-CoV-2 encodes for the viral replication complex, which is essential for viral replication and transcription. Moreover, the L3829F mutation has been found in various non-human hosts like *Felis catus*, *Mesocricetus auratus*, *Mus musculus*, *Neovison vison* and *Odocoileus virginianus*.

Compared to XBB.1.5* and XBB.1.9*, XBB.1.16* variant possesses mutations E180V and T478R in the spike protein. XBB.1.5* and XBB.1.9* carry a T478K mutation, while XBB.1.16* has a T478R mutation. Notably, position 478 lies within the spike protein's binding site to the ACE2 receptor (Figure 4C). However, the lack of a significant difference (*P*-value > 0.05) in ACE2 binding affinity between the T478R and T478K mutations is in line with the literature, which indicates that the spike proteins of XBB.1.16 and XBB.1.5 share similar characteristics in terms of infectivity (86). T478R and T478K mutations show no distinction in terms of antigenicity and immunogenicity. E180V also shows no difference in antigenicity and immunogenicity compared to the reference. XBB.1.16 and XBB.1.5 exhibit similar characteristics in escape ability from humoral immunity (86). Notably, when comparing T478R and T478K mutations, T478R (ΔDDG > 0) enhances the stability of the S protein, while T478K (ΔDDG < 0) reduces it. E180V (ΔDDG > 0) also enhances the stability of the S protein.

The D1746Y mutation is located in the ORF1b gene, which encodes the RNA-dependent RNA polymerase (RdRp), a critical component essential for viral replication (87). Although it is unclear how this mutation affects viral pathogenesis, it has been suggested that the D1746Y mutation increases the stability of the protein (ΔDDG > 0), thereby increasing its replication rate (88). Our results indicate that the D1746Y mutation is undergoing positive selection, which benefits the virus's survival. XBB.1.16 is an XBB.1 sub-lineage (85,89). The sequence carrying the D1746Y mutation accounts for 0.12% of all XBB.1 sequences in the XBB.1 lineage. In XBB.1, the frequency of sequences carrying the D1746Y mutation has significantly increased (Figure 4D). This suggests an adaptation of the D1746Y mutation within the XBB.1 lineage. Within XBB.1, the D1746Y mutation exhibits a robust positive correlation with two other mutations, E180V and D371D (correlation coefficient > 0.6, Holm–Bonferroni corrected *P*-value < 0.05). The E180V mutation is also a characteristic feature of XBB.1.16. Since the spike proteins of XBB.1.16 and XBB.1.5 exhibit similar characteristics in terms of infectivity and escape ability from humoral immunity, the increased fitness of XBB.1.16 might be attributed to the mutations in non-spike proteins (86). The D1746Y mutation may play an essential role in the increased fitness of XBB.1.16.

XBB.1.16* is a newly identified variant that emerged after our data collection. In contrast to XBB.1.5* and XBB.1.9*, all the unique mutations identified in the XBB.1.16* variant are included in the set of 9832 common mutations. The

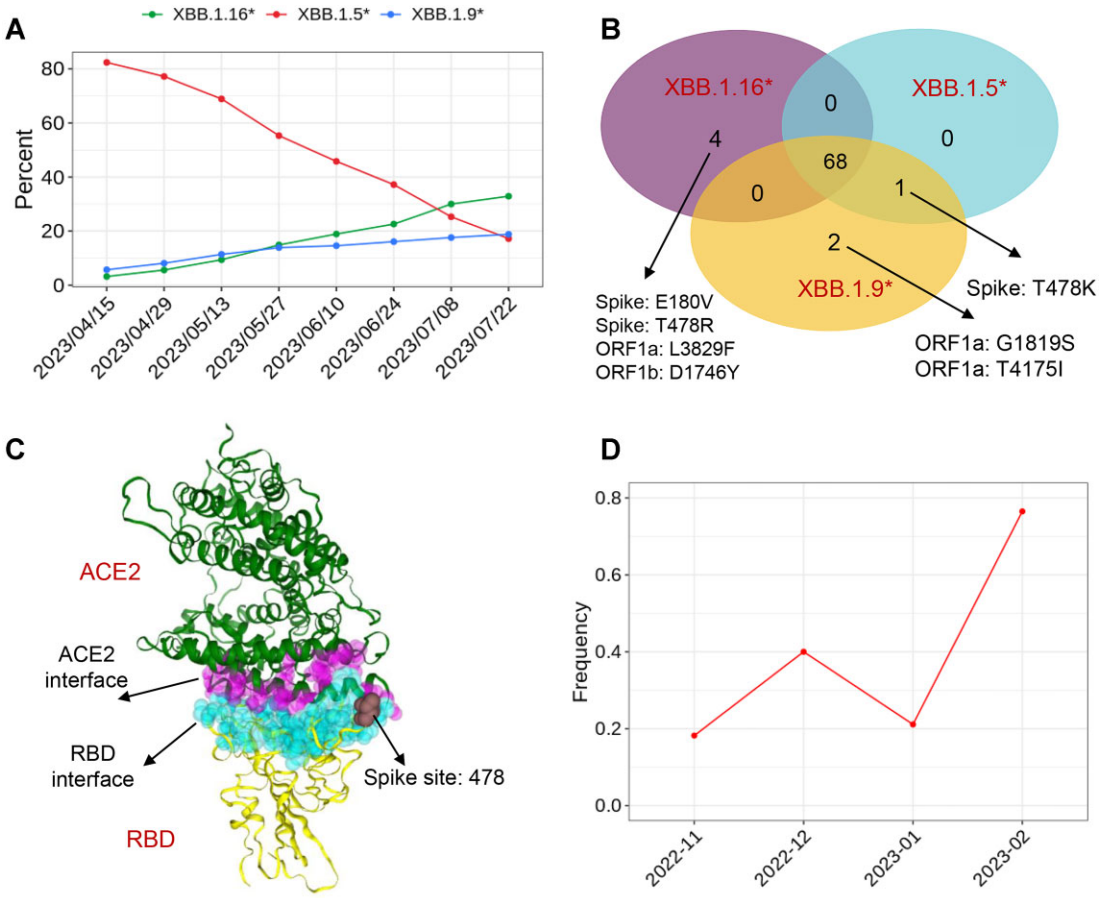


Figure 4. Exploring the latest variant XBB.1.16* using COV2Var. (**A**) Temporal dynamics of sequence counts for XBB.1.5*, XBB.1.16* and XBB.1.19* Variants. (**B**) Mutational differences among XBB.1.5*, XBB.1.16* and XBB.1.19* variants. (**C**) The position T478 in SARS-CoV-2 spike protein. SARS-CoV-2 RBD is shown in yellow, ACE2 in green. The magenta area represents the ACE2 binding interface, while the cyan region signifies the RBD binding interface. (**D**) Frequency changes of sequences carrying D1746Y mutation over time in XBB.1 lineage.

SARS-COV2 has adopted a strategy of survival through a slow pace of accepting new mutations and instead relies on combining mutations (6). By utilizing COV2Var, we can assess the functional impact of mutations in newly emerging variants, aiding in the characterization of these novel variant features.

Discussion

COV2Var is the first and unique database that systematically analyzed over 13 billion SARS-CoV-2 genome sequences and metadata of sequences. We identified 9832 common mutations with competitive advantages. These mutations exhibit greater fitness and spread compared to others, underscoring their advantageous role in evolution and biological significance. Subsequently, we performed a comprehensive functional annotation by integrating sequence and metadata information for a total of 9832 commonly mutated sites. By using COV2Var, we can uncover potential reasons for the notable differences in mutation frequency between the N501Y and N501T mutations in human. For example, the E484K mutation contributes to the competitive advantage of N501Y. N501Y mutation exhibited enhanced RBD affinity for the ACE2 receptor, contributing to increased N501Y transmissibility. The N501T mutation shows greater adaptability to non-human animal hosts. Furthermore, we have employed COV2Var to explore the newly emerged variant XBB.1.16*. Compared to XBB.1.5, XBB.1.16* displays four distinct mutations. The D1746Y mutation may play an essential role increased fitness of XBB.1.16. The L3829F mutation is linked to a higher hospitalization rate. Additionally, the T478R and E180V mutations contribute to enhanced stability in the spike protein. Our database will serve as a unique and comprehensive resource for assessing the impact of SARS-CoV-2 mutations and gaining insights into the characteristics of variants. To maintain COV2Var as a premier mutation annotation database of COVID-19, we will consistently acquire and integrate new data updates into our database.

Data availability

We retrieved genome sequences and metadata from GISAID (https://gisaid.org/). Detailed information about the GISAID data used can be found at https://doi.org/10.55876/gis8.230705yx.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We gratefully acknowledge all data contributors, i.e. the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

Funding

Feng, Yi and Yang were supported by the Center of Excellence-International Collaboration Initiative Grant [139170052]; West China Hospital, Sichuan University and Sichuan Science and Technology Program [2022YFS0228, 2023YFS0200]; Zhao, Wen, Wang and Zhou were supported by NIH [R01GM123037, U01AR069395-01A1, R01CA241930]; NSF [2217515, 2326879]; Dr Kim was supported by NIH [R35GM138184]. Funding for open access charge: Center of Excellence-International Collaboration Initiative Grant [139170052]; West China Hospital, Sichuan University Sichuan Science and Technology Program [2022YFS0228, 2023YFS0200].

Conflict of interest statement

None declared.

References

1. Markov,P.V., Ghafari,M., Beer,M., Lythgoe,K., Simmonds,P., Stilianakis,N.I. and Katzourakis,A. (2023) The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.*, **21**, 361–379.
2. Lauring,A.S. and Hodcroft,E.B. (2021) Genetic variants of SARS-CoV-2-what do they mean? *JAMA*, **325**, 529–531.
3. Woolhouse,M.E.J. and Gowtage-Sequeria,S. (2005) Host range and emerging and reemerging pathogens. *Emerg. Infect. Dis.*, **11**, 1842–1847.
4. Zhang,L., Jackson,C.B., Mou,H., Ojha,A., Peng,H., Quinlan,B.D., Rangarajan,E.S., Pan,A., Vanderheiden,A. and Suthar,M.S. (2020) SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat. Commun.*, **11**, 6013.
5. Xie,X., Liu,Y., Liu,J., Zhang,X., Zou,J., Fontes-Garfias,C.R., Xia,H., Swanson,K.A., Cutler,M. and Cooper,D. (2021) Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera. *Nat. Med.*, **27**, 620–621.
6. Patro,L.P.P., Sathyaseelan,C., Uttamrao,P.P. and Rathinavelan,T. (2021) The evolving proteome of SARS-CoV-2 predominantly uses mutation combination strategy for survival. *Comput. Struct. Biotechnol. J.*, **19**, 3864–3875.
7. Chen,A.T., Altschuler,K., Zhan,S.H., Chan,Y.A. and Deverman,B.E. (2021) COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest. *Elife*, **10**, e63409.
8. Alam,I., Radovanovic,A., Incitti,R., Kamau,A.A., Alarawi,M., Azhar,E.I. and Gojobori,T. (2021) CovMT: an interactive SARS-CoV-2 mutation tracker, with a focus on critical variants. *Lancet Infect. Dis.*, **21**, 602.
9. Singer,J., Gifford,R., Cotten,M. and Robertson,D. (2020) CoV-GLUE: a web application for tracking SARS-CoV-2 genomic variation. Preprints doi: https://doi.org/10.20944/preprints202006.0225.v1, 18 June 2020, preprint: not peer reviewed.
10. Mercatelli,D., Triboli,L., Fornasari,E., Ray,F. and Giorgi,F.M. (2021) Coronapp: a web application to annotate and monitor SARS-CoV-2 mutations. *J. Med. Virol.*, **93**, 3238–3245.
11. Fang,S., Li,K., Shen,J., Liu,S., Liu,J., Yang,L., Hu,C.-D. and Wan,J. (2021) GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences. *Nucleic Acids Res.*, **49**, D706–D714.
12. Gangavarapu,K., Latif,A.A., Mullen,J.L., Alkuzweny,M., Hufbauer,E., Tsueng,G., Haag,E., Zeller,M., Aceves,C.M. and Zaiets,K. (2023) Outbreak. Info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nat. Methods*, **20**, 512–522.
13. Elbe,S. and Buckland-Merrett,G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, **1**, 33–46.
14. O'Toole,Á., Pybus,O.G., Abram,M.E., Kelly,E.J. and Rambaut,A. (2022) Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics*, **23**, 121.
15. Xing,Y., Li,X., Gao,X. and Dong,Q. (2020) MicroGMT: a mutation tracker for SARS-CoV-2 and other microbial genome sequences. *Front. Microbiol.*, **11**, 1502.
16. Sayers,E.W., Cavanaugh,M., Clark,K., Pruitt,K.D., Schoch,C.L., Sherry,S.T. and Karsch-Mizrachi,I. (2022) GenBank. *Nucleic Acids Res.*, **50**, D161–D164.
17. R Core Team (2013) R: A Language and Environment for Statistical Computing.
18. Lucaci,A.G., Zehr,J.D., Shank,S.D., Bouvier,D., Ostrovsky,A., Mei,H., Nekrutenko,A., Martin,D.P. and Kosakovsky Pond,S.L. (2022) RASCL: rapid Assessment of selection in CLades through molecular sequence analysis. *PLoS One*, **17**, e0275623.
19. Pond,S.L.K., Poon,A.F.Y., Velazquez,R., Weaver,S., Hepler,N.L., Murrell,B., Shank,S.D., Magalis,B.R., Bouvier,D., Nekrutenko,A., *et al.* (2020) HyPhy 2.5-A customizable platform for evolutionary hypothesis testing using phylogenies. *Mol. Biol. Evol.*, **37**, 295–299.
20. Pond,S.L.K. and Frost,S.D.W. (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.*, **22**, 1208–1222.
21. Murrell,B., Wertheim,J.O., Moola,S., Weighill,T., Scheffler,K. and Kosakovsky Pond,S.L. (2012) Detecting individual sites subject to episodic diversifying selection. *PLos Genet.*, **8**, e1002764.
22. Murrell,B., Moola,S., Mabona,A., Weighill,T., Sheward,D., Pond,S.L.K. and Scheffler,K. (2013) FUBAR: a fast, unconstrained bayesian AppRoximation for inferring selection. *Mol. Biol. Evol.*, **30**, 1196–1205.
23. Gasteiger,E., Hoogland,C., Gattiker,A., Duvaud,S.e., Wilkins,M.R., Appel,R.D. and Bairoch,A. (2005) In: *Protein Identification and Analysis Tools on the ExPASy Server*. Springer.
24. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.
25. Pejaver,V., Urresti,J., Lugo-Martinez,J., Pagel,K.A., Lin,G.N., Nam,H.-J., Mort,M., Cooper,D.N., Sebat,J. and Iakoucheva,L.M. (2020) Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.*, **11**, 5918.
26. Erdos,G., Pajkos,M. and Dosztanyi,Z. (2021) IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.*, **49**, W297–W303.
27. Doytchinova,I.A. and Flower,D.R. (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinf.*, **8**, 4.
28. Calis,J.J., Maybeno,M., Greenbaum,J.A., Weiskopf,D., De Silva,A.D., Sette,A., Keşmir,C. and Peters,B. (2013) Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.*, **9**, e1003266.
29. Fowler,D.M. and Fields,S. (2014) Deep mutational scanning: a new style of protein science. *Nat. Methods*, **11**, 801–807.
30. Weile,J. and Roth,F.P. (2018) Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. *Hum. Genet.*, **137**, 665–678.

31. Adams,R.M., Mora,T., Walczak,A.M. and Kinney,J.B. (2016) Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *Elife*, **5**, e23156.
32. Starr,T.N., Greaney,A.J., Addetia,A., Hannon,W.W., Choudhary,M.C., Dingens,A.S., Li,J.Z. and Bloom,J.D. (2021) Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science (New York, N.Y.)*, **371**, 850–854.
33. Cao,Y., Jian,F., Wang,J., Yu,Y., Song,W., Yisimayi,A., Wang,J., An,R., Chen,X. and Zhang,N. (2023) Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. *Nature*, **614**, 521–529.
34. Starr,T.N., Czudnochowski,N., Liu,Z., Zatta,F., Park,Y.-J., Addetia,A., Pinto,D., Beltramello,M., Hernandez,P. and Greaney,A.J. (2021) SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. *Nature*, **597**, 97–102.
35. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57**, 289–300.
36. Toyoshima,Y., Nemoto,K., Matsumoto,S., Nakamura,Y. and Kiyotani,K. (2020) SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *J. Hum. Genet.*, **65**, 1075–1082.
37. Becerra-Flores,M. and Cardozo,T. (2020) SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *Int. J. Clin. Pract.*, **74**, e13525.
38. Mahmoudi Gomari,M., Rostami,N., Omidi-Ardali,H. and Arab,S.S. (2022) Insight into molecular characteristics of SARS-CoV-2 spike protein following D614G point mutation, a molecular dynamics study. *J. Biomol. Struct. Dyn.*, **40**, 5634–5642.
39. Fang,S., Liu,S., Shen,J., Lu,A.Z., Wang,A.K.Y., Zhang,Y., Li,K., Liu,J., Yang,L., Hu,C.-D., *et al.* (2021) Updated SARS-CoV-2 single nucleotide variants and mortality association. *J. Med. Virol.*, **93**, 6525–6534.
40. Oulas,A., Zanti,M., Tomazou,M., Zachariou,M., Minadakis,G., Bourdakou,M.M., Pavlidis,P. and Spyrou,G.M. (2021) Generalized linear models provide a measure of virulence for specific mutations in SARS-CoV-2 strains. *PLoS One*, **16**, e0238665.
41. Wang,R., Chen,J., Gao,K., Hozumi,Y., Yin,C. and Wei,G.-W. (2021) Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. *Commun. Biol.*, **4**, 228.
42. Conti,P. and Younes,A. (2020) Coronavirus COV-19/SARS-CoV-2 affects women less than men: clinical response to viral infection. *J. Biol. Regul. Homeost. Agents*, **34**, 339–343.
43. Promislow,D.E.L. (2020) A geroscience perspective on COVID-19 mortality. *J. Gerontol. A Biol. Sci. Med. Sci.*, **75**, E30–E33.
44. Yuan,M., Wu,N.C., Zhu,X., Lee,C.-C.D., So,R.T., Lv,H., Mok,C.K. and Wilson,I.A. (2020) A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science*, **368**, 630–633.
45. Martin,D.P., Weaver,S., Tegally,H., San,J.E., Shank,S.D., Wilkinson,E., Lucaci,A.G., Giandhari,J., Naidoo,S. and Pillay,Y. (2021) The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell*, **184**, 5189–5200.
46. Al-Khatib,H.A., Smatti,M.K., Ali,F.H., Zedan,H.T., Thomas,S., Ahmed,M.N., El-kahlout,R.A., Al Bader,M.A., Elgakhlab,D., Coyle,P.V., *et al.* (2022) Comparative analysis of within-host diversity among vaccinated COVID-19 patients infected with different SARS-CoV-2 variants. *Iscience*, **25**, 105438.
47. Oreshkova,N., Molenaar,R.J., Vreman,S., Harders,F., Munnink,B.B.O., Hakze-van Der Honing,R.W., Gerhards,N., Tolsma,P., Bouwstra,R., Sikkema,R.S., *et al.* (2020) SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. *Eurosurveillance*, **25**, 2–8.
48. Hobbs,E.C. and Reid,T.J. (2021) Animals and SARS-CoV-2: species susceptibility and viral transmission in experimental and natural conditions, and the potential implications for community transmission. *Transbound. Emerg. Dis.*, **68**, 1850–1867.
49. Yen,H.-L., Sit,T.H., Brackman,C.J., Chuk,S.S., Cheng,S., Gu,H., Chang,L.D., Krishnan,P., Ng,D.Y. and Liu,G.Y. (2022) Transmission of SARS-CoV-2 (variant delta) from pet hamsters to humans and onward human propagation of the adapted strain: a case study. *Lancet*, **399**, 1070–1078.
50. Ren,W., Lan,J., Ju,X., Gong,M., Long,Q., Zhu,Z., Yu,Y., Wu,J., Zhong,J. and Zhang,R. (2021) Mutation Y453F in the spike protein of SARS-CoV-2 enhances interaction with the mink ACE2 receptor for host adaption. *PLoS Pathog.*, **17**, e1010053.
51. Zhou,J., Peacock,T.P., Brown,J.C., Goldhill,D.H., Elrefaey,A.M., Penrice-Randal,R., Cowton,V.M., De Lorenzo,G., Furnon,W. and Harvey,W.T. (2022) Mutations that adapt SARS-CoV-2 to mink or ferret do not increase fitness in the human airway. *Cell Rep.*, **38**, 110344.
52. Tenchov,R. and Zhou,Q.A. (2022) Intrinsically disordered Proteins: perspective on COVID-19 infection and drug discovery. *Acs Infectious Diseases*, **8**, 422–432.
53. Anjum,F., Mohammad,T., Asrani,P., Shafie,A., Singh,S., Yadav,D.K., Uversky,V.N. and Hassan,M.I. (2022) Identification of intrinsically disorder regions in non-structural proteins of SARS-CoV-2: new insights into drug and vaccine resistance. *Mol. Cell. Biochem.*, **477**, 1607–1619.
54. Quaglia,F., Salladini,E., Carraro,M., Minervini,G., Tosatto,S.C.E. and Le Mercier,P. (2022) SARS-CoV-2 variants preferentially emerge at intrinsically disordered protein sites helping immune evasion. *FEBS J.*, **289**, 4240–4250.
55. Bai,Z., Cao,Y., Liu,W. and Li,J. (2021) The SARS-CoV-2 nucleocapsid protein and its role in viral structure, biological functions, and a potential target for drug or vaccine mitigation. *Viruses*, **13**, 1115.
56. Azad,G.K. (2021) Identification and molecular characterization of mutations in nucleocapsid phosphoprotein of SARS-CoV-2. *PeerJ*, **9**, e10666.
57. Kumar,S., Thambiraja,T.S., Karuppanan,K. and Subramaniam,G. (2022) Omicron and delta variant of SARS-CoV-2: a comparative computational study of spike protein. *J. Med. Virol.*, **94**, 1641–1649.
58. Pondé,R. (2022) Physicochemical effect of the N501Y, E484K/Q, K417N/T, L452R and T478K mutations on the SARS-CoV-2 spike protein RBD and its influence on agent fitness and on attributes developed by emerging variants of concern. *Virology*, **572**, 44–54.
59. Prompetchara,E., Kettoy,C. and Palaga,T. (2020) Immune responses in COVID-19 and potential vaccines: lessons learned from SARS and MERS epidemic. *Asian Pac. J. Allergy Immunol.*, **38**, 1–9.
60. Li,W., Li,L., Sun,T., He,Y., Liu,G., Xiao,Z., Fan,Y. and Zhang,J. (2020) Spike protein-based epitopes predicted against SARS-CoV-2 through literature mining. *Med. Novel Technol. Devices*, **8**, 100048–100048.
61. Kadam,S.B., Sukhramani,G.S., Bishnoi,P., Pable,A.A. and Barvkar,V.T. (2021) SARS-CoV-2, the pandemic coronavirus: molecular and structural insights. *J. Basic Microbiol.*, **61**, 180–202.
62. Yang,Y., Zhang,Y., Qu,Y., Zhang,C., Liu,X.-W., Zhao,M., Mu,Y. and Li,W. (2021) Key residues of the receptor binding domain in the spike protein of SARS-CoV-2 mediating the interactions with ACE2: a molecular dynamics study. *Nanoscale*, **13**, 9364–9370.
63. Wrapp,D., Wang,N., Corbett,K.S., Goldsmith,J.A., Hsieh,C.-L., Abiona,O., Graham,B.S. and McLellan,J.S. (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, **367**, 1260–1263.
64. Sanches,P.R., Charlie-Silva,I., Braz,H.L., Bittar,C., Calmon,M.F., Rahal,P. and Cilli,E.M. (2021) Recent advances in SARS-CoV-2 spike protein and RBD mutations comparison between new variants alpha (B. 1.1. 7, United Kingdom), Beta (B. 1.351, South Africa), Gamma (P. 1, Brazil) and delta (B. 1.617. 2, India). *J. Virus Eradic.*, **7**, 100054.
65. Piccoli,L., Park,Y.-J., Tortorici,M.A., Czudnochowski,N., Walls,A.C., Beltramello,M., Silacci-Fregni,C., Pinto,D., Rosen,L.E. and Bowen,J.E. (2020) Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding

domain by structure-guided high-resolution serology. *Cell*, **183**, 1024–1042.
66. Liu,L., Iketani,S., Guo,Y., Chan,J.F.-W., Wang,M., Liu,L., Luo,Y., Chu,H., Huang,Y. and Nair,M.S. (2022) Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature*, **602**, 676–681.
67. Wang,Q., Guo,Y., Iketani,S., Nair,M.S., Li,Z., Mohri,H., Wang,M., Yu,J., Bowen,A.D. and Chang,J.Y. (2022) Antibody evasion by SARS-CoV-2 Omicron subvariants BA. 2.12. 1, BA. 4 and BA. 5. *Nature*, **608**, 603–608.
68. Wang,Q., Iketani,S., Li,Z., Guo,Y., Yeh,A.Y., Liu,M., Yu,J., Sheng,Z., Huang,Y. and Liu,L. (2022) Antigenic characterization of the SARS-CoV-2 Omicron subvariant BA. 2.75. *Cell Host Microbe*, **30**, 1512–1517.
69. Baum,A., Fulton,B.O., Wloga,E., Copin,R., Pascal,K.E., Russo,V., Giordano,S., Lanza,K., Negron,N. and Ni,M. (2020) Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science*, **369**, 1014–1018.
70. Andreano,E., Piccini,G., Licastro,D., Casalino,L., Johnson,N.V., Paciello,I., Dal Monego,S., Pantano,E., Manganaro,N. and Manenti,A. (2021) SARS-CoV-2 escape from a highly neutralizing COVID-19 convalescent plasma. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2103154118.
71. Weisblum,Y., Schmidt,F., Zhang,F., DaSilva,J., Poston,D., Lorenzi,J.C., Muecksch,F., Rutkowska,M., Hoffmann,H.-H. and Michailidis,E. (2020) Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *Elife*, **9**, e61312.
72. Liu,Z., VanBlargan,L.A., Bloyet,L.-M., Rothlauf,P.W., Chen,R.E., Stumpf,S., Zhao,H., Errico,J.M., Theel,E.S. and Liebeskind,M.J. (2021) Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host Microbes Infect.*, **29**, 477–488.
73. Li,Q., Wu,J., Nie,J., Zhang,L., Hao,H., Liu,S., Zhao,C., Zhang,Q., Liu,H. and Nie,L. (2020) The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell*, **182**, 1284–1294.
74. Rahman,M.S., Islam,M.R., Ul Alam,A.S.M.R., Islam,I., Hoque,M.N., Akter,S., Rahaman,M.M., Sultana,M. and Hossain,M.A. (2021) Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences. *J. Med. Virol.*, **93**, 2177–2195.
75. Mahmoudi Gomari,M., Rostami,N., Omidi-Ardali,H. and Arab,S.S. (2022) Insight into molecular characteristics of SARS-CoV-2 spike protein following D614G point mutation, a molecular dynamics study. *J. Biomol. Struct.*, **40**, 5634–5642.
76. Periwal,N., Rathod,S.B., Sarma,S., Johar,G.S., Jain,A., Barnwal,R.P., Srivastava,K.R., Kaur,B., Arora,P. and Sood,V. (2022) Time series analysis of SARS-CoV-2 genomes and correlations among highly prevalent mutations. *Microbiol. Spectrum*, **10**, e0121922.
77. Walker,A.S., Vihta,K.-D., Gethings,O., Pritchard,E., Jones,J., House,T., Bell,I., Bell,J.I., Newton,J.N., Farrar,J., *et al.* (2021) Tracking the emergence of SARS-CoV-2 alpha variant in the United Kingdom. *N. Engl. J. Med.*, **385**, 2582–2585.
78. Mohseni Afshar,Z., Babazadeh,A., Janbakhsh,A., Mansouri,F., Sio,T.T., Sullman,M.J., Carson-Chahhoud,K., Hosseinzadeh,R., Barary,M. and Ebrahimpour,S. (2022) Coronavirus disease 2019 (Covid-19) vaccination recommendations in special populations and patients with existing comorbidities. *Rev. Med. Virol.*, **32**, e2309.
79. Fisman,D.N. and Tuite,A.R. (2021) Evaluation of the relative virulence of novel SARS-CoV-2 variants: a retrospective cohort study in Ontario, Canada. *Can. Med. Assoc. J.*, **193**, E1619–E1625.
80. Hanafusa,M., Kuramochi,J., Ishihara,K., Honda,M., Nawa,N. and Fujiwara,T. (2021) Clinical characteristics of patients with SARS-CoV-2 N501Y variants in general practitioner clinic in Japan. *J. Clin. Med.*, **10**, 5865.
81. Khare,S., Gurry,C., Freitas,L., Schultz,M.B., Bach,G., Diallo,A., Akite,N., Ho,J., Lee,R.T.C., Yeo,W., *et al.* (2021) GISAID's role in pandemic response. *China Cdc Weekly*, **3**, 1049–1051.
82. Yeh,T.-Y. and Contreras,G.P. (2021) Viral transmission and evolution dynamics of SARS-CoV-2 in shipboard quarantine. *Bull. World Health Organ.*, **99**, 486–495.
83. Stolp,B., Stern,M., Ambiel,I., Hofmann,K., Morath,K., Gallucci,L., Cortese,M., Bartenschlager,R., Ruggieri,A. and Graw,F. (2022) SARS-CoV-2 variants of concern display enhanced intrinsic pathogenic properties and expanded organ tropism in mouse models. *Cell Rep.*, **38**, 110387.
84. Carlin,A.F., Clark,A.E., Chaillon,A., Garretson,A.F., Bray,W., Porrachia,M., Santos,A.T., Rana,T.M. and Smith,D.M. (2023) Virologic and immunologic characterization of coronavirus disease 2019 recrudescence after nirmatrelvir/ritonavir treatment. *Clin. Infect. Dis.*, **76**, e530–e532.
85. Karyakarte,R.P., Das,R., Rajmane,M.V., Dudhate,S., Agarasen,J., Pillai,P., Chandankhede,P.M., Labhshetwar,R.S., Gadiyal,Y. and Kulkarni,P.P. (2023) Chasing SARS-CoV-2 XBB. 1.16 recombinant lineage in India and the clinical profile of XBB. 1.16 cases in Maharashtra, India. *Cureus*, **15**, e39816.
86. Genotype to Phenotype Japan (G2P-Japan) Consortium, Yamasoba,D., Uriu,K., Plianchaisuk,A., Kosugi,Y., Pan,L., Zahradnik,J., Ito,J. and Sato,K.(2023) Virological characteristics of the SARS-CoV-2 omicron XBB.1.16 variant. *Lancet Infect. Dis.*, **23**, 655–656.
87. Aftab,S.O., Ghouri,M.Z., Masood,M.U., Haider,Z., Khan,Z., Ahmad,A. and Munawar,N. (2020) Analysis of SARS-CoV-2 RNA-dependent RNA polymerase as a potential therapeutic drug target using a computational approach. *J. Transl. Med.*, **18**, 275.
88. Ramaiah,A., Khubbar,M., Akinyemi,K., Bauer,A., Carranza,F., Weiner,J., Bhattacharyya,S., Payne,D. and Balakrishnan,N. (2023) Genomic surveillance reveals the rapid expansion of the XBB lineage among circulating SARS-CoV-2 Omicron lineages in Southeastern Wisconsin, USA. *Viruses*, **15**, 1940.
89. John,O.O., Olabode,O.N., Lucero-Prisno,D.E. III, Adebimpe,O.T. and Singh,A. (2023) XBB. 1.16 omicron subvariant rise to a variant of interest: implications for global alertness and preparedness. *J. Taibah Univ. Med. Sci.*, **18**, 1285–1287.

Received: August 15, 2023. **Revised:** September 29, 2023. **Editorial Decision:** October 12, 2023. **Accepted:** October 16, 2023