Dynamic Expedition of Leading Mutations in SARS-CoV-2 Spike Glycoproteins

Zhouyi He*,[1,2] Muhammad Hasan*,[1,2] Mengqi Jia*,[1] Kathiresan Natarajan,[3]
Shan Qi Yap,[1] Feng Zhou,[1] Hailei Su†,[4] Kaicheng Zhu†,[1] and Haibin Su†[1,2]

[1] *Department of Chemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China*

[2] *Hong Kong Branch of the Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China*

[3] *Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram - 695014, Kerala, India*

[4] *Bengbu Hospital of Traditional Chinese Medicine, 4339 Huai-shang Road, Anhui 233080, China*

(Dated: December 29, 2021)

During the ongoing CoVID-19 epidemic, the continuous genomic evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been generating new variants with enhanced transmissibility and immune escape. Being one key target of antibodies, mutations of the spike glycoprotein play a vital role in the trajectory of virus evasion. Here, we present a time-resolved statistical method, dynamic expedition of leading mutations (deLemus), to analyze the evolution dynamics of the spike protein. Together with analysis on single amino-acid polymorphism (SAP), we proposed one $\boldsymbol{L}$-index to quantify the mutation strength of each amino acid for unravelling mutation pattern of spike glycoprotein. The sites of interest (SOI) with high $\boldsymbol{L}$-index hold great promise to detect potential signal of emergent variants.

Introduction

The advent of coronavirus disease 2019 (CoVID-19) in Dec. 2019, by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and its subsequent global spread have elevated it a major global threat.[1] The epidemic not only has a significant impact on the economic and social activities, but also has taken away millions of lives.[2] Therefore, efforts have been taken to mitigate the effects of this massive outbreak such as implementing mass vaccination and enforcing lockdowns. The genome size of coronaviruse ranges from approximately 26 to 32 kilobases, among the largest of RNA viruses.[3] Together with its translational frameshifting mechanism in open reading frame 1a and 1b, coronavirus exhibits diversity in encoded proteins and corresponding functions. SARS-CoV-2 has $3'$ - $5'$ exonuclease proofreading of nonstructural protein 14 that has a lower substitution rate than other RNA viruses.[4–6] However, the proofreading function of SARS-CoV-2 cannot repair deletions resulting in viruses with changed protein sequences over time. The SARS-COV-2 genome mutation rate is in the order of 10^{-3} nucleotide substitution per nucleotide site per year.[7] Recombination between variants may also add to the virulence and severity of SARS-CoV-2.[8] The potential recombination region is found inside the spike glycoprotein's receptor-binding domain (RBD), suggesting that the virus might evolve in a directed manner. The emergence of deletions can provide a new pathway for the evolution of coronavirus, leading to highly transmissible variants.

As the pandemic rages on, the virus evolves, and the genetic diversity increases. Evolution happens when there is a lot of genetic diversity, which then meets a selection pressure. Continuous genomic surveillance is possible due to high-throughput worldwide genome sequencing and data sharing via GISAID.[9] Over a hundred additional SARS-COV-2 lineages have emerged in the last two years.[10] Several variants including α (B.1.1.7), β (B.1.351), γ (P.1), δ (B.1.617.2) and o (B.1.1.529) were declared as VoC by the World Health Organization(WHO),[11] as they caused resurgence of CoVID-19. Additionally, mutations in these viruses could make it more resistant to vaccines. The relentless mutation of this coronavirus complicates vaccine and drug research, yet the complexity of biological science and cognitive biases prevent quantitative capture of viral evolution dynamics.[12] Across the genome landscape of SARS-COV-2, the spike glycoprotein coding region is the most active plateau as seen by residue diversity, which is the key mediator of viral entry and fusion into target cells. Spike glycoprotein is a trimeric type I viral fusion protein that binds virus to the Angiotensin-converting enzyme 2 (ACE2) receptor on host cells.[13–16] Spike proteins are glycosylated with N-

[*] These authors contribute equally.
[†] Corresponding authors.

linked glycans, regulating host proteases and neutralizing antibodies accessibility.[17] The spike glycoprotein monomer is composed of 22 N-linked glycosylation sequons.[18] The N-terminal region of spike glycoprotein facilitates ACE2 binding, while the C-terminal region mediates fusion of viral and cellular membranes.[19] N-terminal S1 subunit of spike glycoprotein has an N-terminal domain (NTD) and a receptor-binding domain (RBD) crucial for virus–receptor attachment and a C-terminal S2 subunit for virus–cell membrane fusion. The spike glycoprotein interacts with host cell membrane receptor ACE2 through S1-RBD-interaction-triggering conformational changes in the S2 subunit, resulting in virus fusion and cell entry.[16,20,21] SARS-CoV-2 spike glycoprotein appears to be more mutagenic, most likely due to its critical function in accessing host cell and modifying infectivity. The active mutation in spike glycoproteins that results in immunological escape or increased fitness is a major concern.[22,23]

Despite the significant efforts contributed by the global collaborations in clinical and laboratory investigations, the understanding of the spike glycoprotein evolution is still in its infancy. The complete sequence space of spike glycoprotein consists of $> 20^{1000}$ different sequences formed by all possible amino acid combinations of around 1273 residues. Even the most recent evolutionary trajectory encompasses a very small fraction of the whole sequence space with less than a few millions ($\sim 10^6$) of distinct sequences. Although nearly all of the spike glycoprotein residues have been found with point mutations, the SAP at distinct locations diverges and the co-mutations of different residues uncommon at the whole sequence level. In other words, the majority of the possible sequences remain unexplored, as the majority of the spike glycoprotein sequences already found are concentrated in a small region of the sequence space. This demonstrates that the evolutionary process of spike glycoprotein is very dynamical, and it reveals the significance of understanding how spike glycoprotein navigates the sequence space. In this work, a method named as dynamical expedition of leading mutations (deLemus) is developed to quantitatively characterize the robust characteristics of the evolutionary dynamics of spike glycoprotein. Variations in both mutation rate and single amino acid polymorphism (SAP) of the spike glycoprotein provide valuable information of SARS-CoV-2 variants' temporal evolution. The mutation patterns of spike glycoproteins are explicitly depicted from the whole-sequence to the single-residue level. Particularly, the amino acid SOI of spike glycoprotein are identified through quantitative ranking by $\boldsymbol{L}$-index. The reported SOI in this work include those confirmed by mutations from previous and current VoC/VoI, and outlined ones are possible mutations predicted to be in the forthcoming variants by the deLemus method.

Methodology

Detecting dynamic patterns from big-data sets is always a major challenge in data analysis. In this work, we propose one method, dubbed as deLemus, to investigate the evolution dynamics of the spike glycoprotein at the amino acid (a.a.) sequence level, which is composed of the following three steps.

Data Collection. A collection of total 3,828,746 a.a. sequences of the SARS-CoV-2 spike glycoprotein were downloaded from GISAID hCoV-19 database on Sep 27th 2021.[9] We use EPI_ISL_402124 as the reference sequence of spike glycoprotein.[1,24] As there exist substantial amount of repeated sequences in the original data, all the identical sequences are removed to keep the non-degenerate sequences of spike glycoprotein for the further analysis of sequence mutation. Overall, 147,883 non-degenerate sequences are collected from the total 3,828,746 reported sequences from Dec. 2019 to Aug. 2021.

Alignment. The sequences submitted within each month are grouped together. The multiple sequence alignment for individual group was conducted consecutively using Clustal Omega.[25]. For the aligned spike glycoprotein sequences, every sequence was compared with the reference sequence to check the a.a. substitution or deletion at each a.a. site. This yielded the number of mutated a.a. site, $\boldsymbol{n}$, of every sequence, and the number of sequences, $P(\boldsymbol{n})$, was counted for the given $\boldsymbol{n}$ at each month. Subsequently, the mutation rate Ξ in the unit of $seq^{-1}mo^{-1}$ was calculated by the total number of mutations per sequence per month. In addition, the number of SAP at the jth a.a. site in the tth month, $\boldsymbol{s}_j(t)$, was computed. The number of a.a. sites, $N(\boldsymbol{s})$, was counted at given $\boldsymbol{s}$ monthly. Poisson distribution was observed to give the average SAP numbers $\overline{\boldsymbol{s}}$ at each month.

$\boldsymbol{L}$-index. For each month, one $m \times n$ matrix $\boldsymbol{H}(t)$ is constructed based on the multiple sequences alignment data, where m is the number of the non-degenerate sequences displayed in a specific month, and n is the length of the a.a. sequence of spike protein. In other words, each row represents one non-degenerate sequence from that month, and each column corresponds to one residue in spike glycoprotein. For the ith sequence, if the jth residue is changed, the corresponding matrix component is set to 1, i.e., $\boldsymbol{H}_{ij}(t) = 1$, otherwise it is zero. Then we factorized the mutation

matrices $\boldsymbol{H}(t)$ with singular value decomposition (SVD)[26] as follows,

$$\boldsymbol{H}(t) = \boldsymbol{P}(t) \cdot \boldsymbol{\Sigma}(t) \cdot \boldsymbol{N}(t) \tag{1}$$

where $\boldsymbol{P}(t)$ is a $m \times m$ matrix, m is the number of non-degenerate sequences(# seqs) shown in the tth month, $\boldsymbol{N}(t)$ is a $n \times n$ matrix, and n is the length of spike glycoprotein sequence. (see Supplemental Material for details). From the monthly $\sigma_i(t)$ and $\boldsymbol{N}_{ij}(t)$, we collected the top 4 leading sets of mutations and proposed $\boldsymbol{L}$-index to quantify the mutation strength of each jth site for each month,

$$\boldsymbol{L}_j(t) = \boldsymbol{s}_j(t) \cdot \sqrt{\sum_{i=1}^{4} \boldsymbol{N}_{ij}^2(t)} \tag{2}$$

The a.a. sites were ranked according to their $\boldsymbol{L}_j(t)$ index for each month, and top ranked ones were identified as SOI of each month from Jan. 2020 to Aug. 2021.

Results and Discussions

General Features of Mutations in SARS-CoV-2 Spike Glycoproteins.

While majority of the genetic variation is considered neutral, sometimes a single amino-acid change can have a profound effect on the function of protein, and which is of vital significance.[27] single amino-acid change is also referred as SAP, which are usually caused by non-synonymous single-nucleotide polymorphism. As mutations in a.a. level could not reflect the direct impact SARS-CoV-2 than that in genome level, and spike glycoprotein influences the antigenic change the most, our objective is the a.a. mutations in spike glycoprotein.

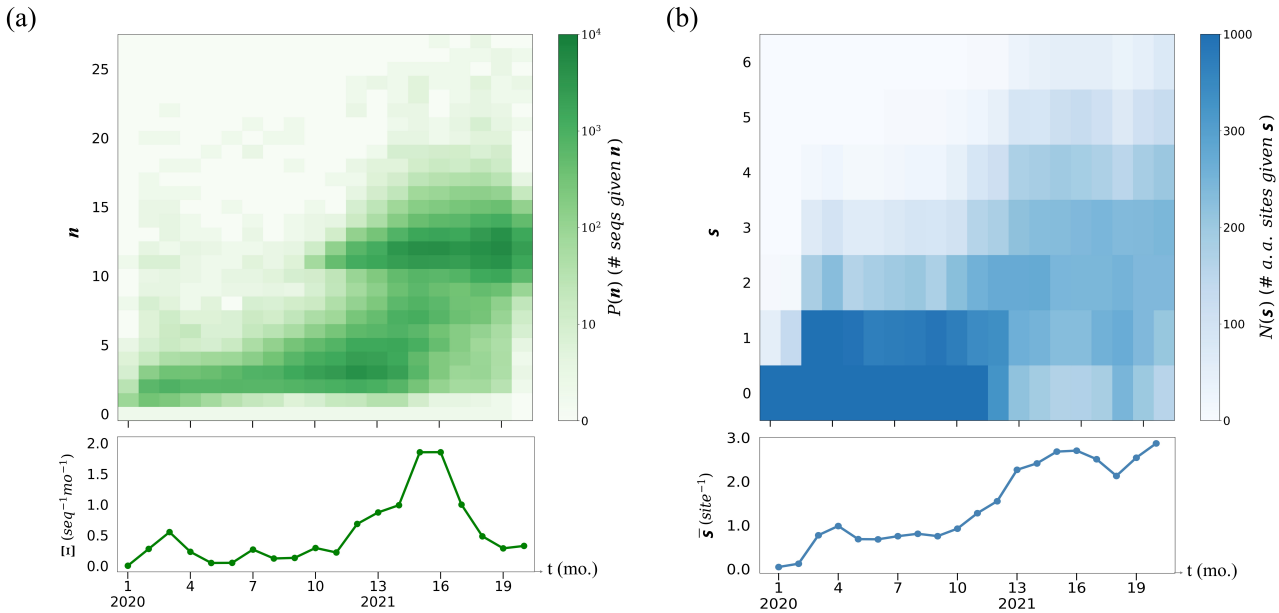


Fig. 1. (a) Time-resolved evolution of the single-residue mutations in the spike protein sequences. The top heat map denotes the time-dependent distribution of n, the number of the mutations per sequence. The bottom curve shows the time course of average mutation rate Ξ. (b) Time-resolved evolution of SAP. The top heat map denotes the time-dependent distribution of s, the number of different a.a. substitutions at every site. The bottom curve shows the time course of average substitution number $\bar{s}$.

As seen in Fig. 1(a), two distinct peaks are shown in the distribution of $P(\boldsymbol{n})$.[1,24] In the early stage prior to the massive spread of the VoC/VoI, most of the sequences exhibit just two or three mutations, including D614G and N501Y. In Oct. 2020, n jumps to 11 drmatically, which is, coincidentally, at the emergence time of α and β variants, containing 10 and 11 mutations in spike glycoprotein, respectively (Fig. S3). After Jan. 2021, more variants started

to appear and their $\boldsymbol{n}$ were all close to 12. In nutshell, change of peak location represent the new evolutionary status: the dominance of wild type strains are being replaced by variants with clustered mutations, such as α and variants. $\bar{\boldsymbol{n}}$ have reflected the ensemble average deviation compared to the reference sequence (Fig. S4), and the increments of $\bar{\boldsymbol{n}}$ in each month are difined as mutation rate (Fig. 1(b)). $\Xi_{t_k} = [\sum_{ij} \boldsymbol{H}_{ij}(t_k) - \sum_{ij} \boldsymbol{H}_{ij}(t_{k-1})]/[(t_k - t_{k-1}) \cdot m(t_k)]$, where $H_{i,j}(t_k)$ is the mutation matrix and $m(t_k)$ is number of sequence in the kth month. The curve of Ξ_{t_k} is shaped coherently by both natural drift and natural selection The increase between Dec. 2020 and Jun. 2021 is consistent with the spread of VoC/VoI (mostly α and δ variants). Ξ achieved maximum around Mar. and Apr. 2021, which correspond to the transition state between early stage strains and α variant, as showed in Fig. S1(a). The relatively low Ξ value indicates the dominance of proximate early strains, from Jan. 2020 to Nov. 2020, and the prevalence of δ variant since Jun. 2021.

After knowing the ensemble property of individual sequences, however, mutations at some key sites could make a big difference. Thus, we defined different types of SAPs in given site j as $\boldsymbol{s_j}$, which reflects the mutation activity of site j. The dynamics of $N(\boldsymbol{s})$, which specifies the total count of sites containing $\boldsymbol{s}$ distinct SAPs (Fig. 1(b) top panel and Fig. S2(a)), shows the diversity of full a.a. sites are in general increasing. Distribution of $\boldsymbol{s_j}$ of different monthes all fit well with Poisson distribution and $\bar{\boldsymbol{s}}$ (mean of number of SAPs) is retrieved afterwards (Fig. 1(b) bottom panel). Poisson distribution is given by $N(\boldsymbol{s}) = N_0 e^{-\bar{\boldsymbol{s}}} \bar{\boldsymbol{s}}^{\boldsymbol{s}} / \boldsymbol{s}!$, where N_0 is the length of spike sequence 1273 a.a. and and f is the mean value of $\boldsymbol{s}$. As can be seen in Fig. 1(b), during the early stage of pandemic, the majority of the sites only has a very low $\boldsymbol{s}$ around 1. Once more, with the rapid spread of VoC/VoI, $\bar{\boldsymbol{s}}$ increased dramatically and remained at a high level ever since, which is indicative of the ongoing adaptive diversification. We have also found that the majority of sites included as part of VoC/VoI exhibit high $\boldsymbol{s}$, indicating that active mutations could lead to the selection advantage by a cluster of mutations, as each SAP at each single site is likely be optimized. Although the emergence of new SAPs are inevitable consequences of the uncertainty in genome replication processes, $\bar{\boldsymbol{s}}$ could be indicative of the virus diversity at given time window.

Our results clearly demonstrate that spike glycoproteins are undergoing active adaptive diversification based on ensemble properties $\boldsymbol{n}$ and $\boldsymbol{s}$, which represent the drift in the sequence space and single site diversity represented, respectively. Emergence of SARS-CoV-2 variants with novel spike glycoprotein mutations have influenced the epidemiological and clinical aspects of COVID-19 pandemic. To implement a surveillance for evolution of spike glycoprotein, we need methods to analyze time resolved sequence data with a.a. site resolution, such as deLemus, which is developed to meet this need to quantify the mutation strength of each a.a. in this work.

$\boldsymbol{L}$-indexed Mutated Sites in SARS-CoV-2 Spike Glycoproteins

Although the mechanisms by which the key mutations are favored and the formidable variants are established in nature remain unclear, the temporal emergence of leading mutations offers clues to evolutionary trajectory of SARS-CoV-2. The proposed $\boldsymbol{L}$-index evaluates the mutation strength of each a.a., which yields the top ranked ones as our SOI (Table.S2). Interestingly, pre-peaks of key mutations belonging to VoC/VoI are already presented as shown in Fig. S1(a). For instance, the earliest signal arising in Feb. 2020 located at a.a. site 614, indicating the dominating D614G mutation. Then, mutations of α variant are captured as well as mutations from other significant VoC/VoI. We found that NTD showed a very fast response in terms of mutations. For example, $\Delta 69 - 70$ was detected mutations before the emergence of α variant (Fig. S1(b)) and we can also see the coincident signals in RDRs.[28] After Dec. 2020, the identification of five rapidly expanding virus lineages, designated as VoC, has ushered in a new stage of the pandemic. From Fig. S1(b), we can already see the leading peaks of α variant, as well as β and γ variants, which share several mutations with each other particularly in key regions of the spike glycoprotein involved in recognizing the host-cells ACE2 receptor. Another novel leading mutation L452R started to rise around two months before the classification of δ variant, which was reported to increases spike stability, viral infectivity, viral fusogenicity, and thereby promoted viral replication.[29] Once δ variant arrived in the United Kingdom, this variant spread at a unprecedented pace, which was estimated about 60% more transmissible than α variant. Nevertheless, a novel variant, namely o, is expected to be the next dominant variant due to enhanced transmissibility. Despite that additional 50k sequences from Sep. 2021 onwards are still being processed, a large number of spike mutations of this striking variant have already appeared in our preliminary results, which is claimed to be responsible for the escape of majority of existing neutralizing antibodies, posing a serious threat to the development of vaccines and therapies.[30–32] Those key mutations are the crucial contributors for the race of fitness and provides insight[33,34] into how SARS-CoV-2 is changing and how the pandemic will unfold in the future.[35]

Spike NTD mutations. NTD is a significant target for neutralising monoclonal antibodies (mAbs).[36–40] Several

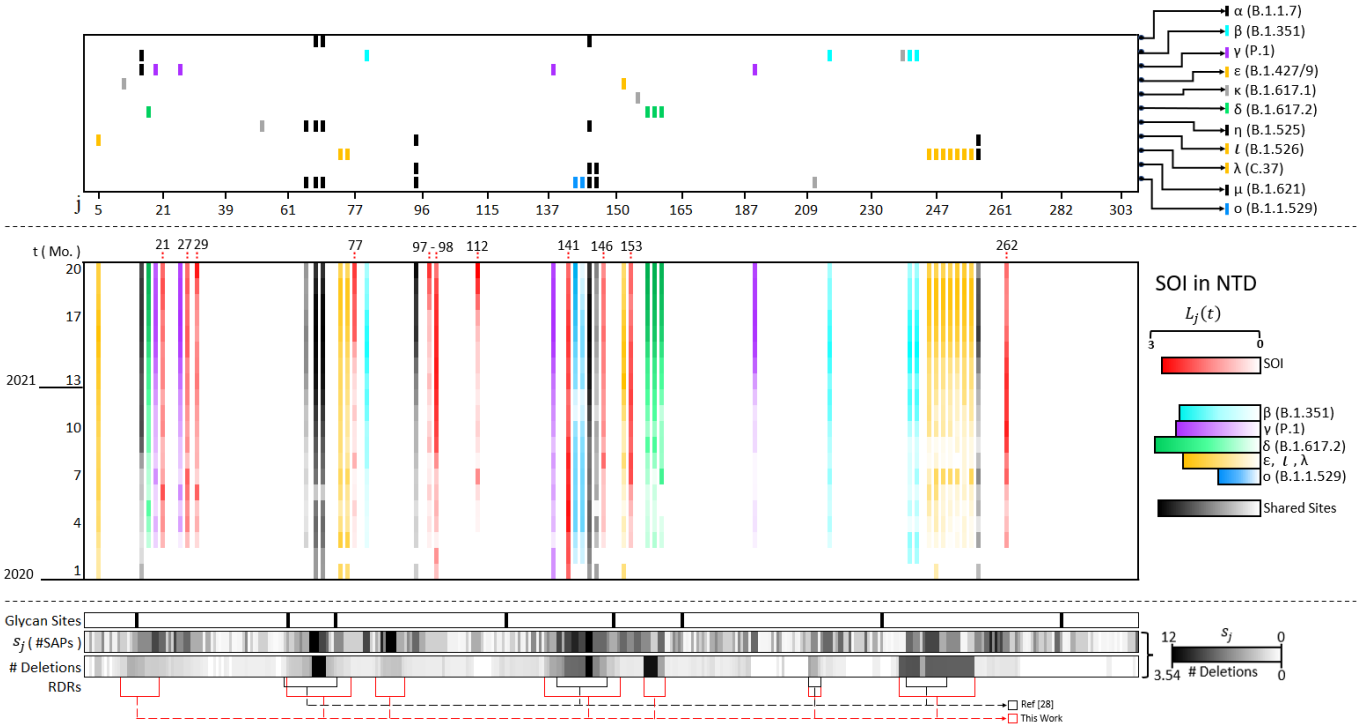


Fig. 2. Top panel: NTD mutations of the spike protein of VoC/VoI defined by WHO. Sites shared by more than one VoC/VoI are colored in black. Middle panel: Dynamic expedition of SOI in NTD. Outlined emergent mutated sites are highlighted by a.a. site indices above. Bottom panel: Glycan sites, s_j (substitution number of a.a. at site j) and deletion occurrence are the mean values from the sequences from Mar. to Aug. 2021. RDRs reported in ref[28] are in black, and RDRs identified in this work are in red.

mAbs directed against NTD preferentially inhibit ACE2-independent cell.[41] This directs to one evolutionary path towards immune evasion leading to high mutation rate in NTD. In addition to a.a. substitution mutations, a substantial number of a.a. deletion mutations connected to immunological escape has been found in NTD.[28] In Fig. 2 Mutations from different VoC/VoI are marked in the top panel, and deletion occurrence as well as RDRs are potted in the bottom panel. Temporal emergence of SOI in NTD are shown in Fig. 2 middle panel with marking scheme based on corresponding variants, which shows strengthened signal as evolution continues. Four previously identified RDRs are also incluede in SOI:[28] $\Delta 69-70$ (N2 loop), $\Delta 141-144$ and $\Delta 146$ (N3 loop), $\Delta 210$ (between N4 and N5 loops), and $\Delta 243-244$ (N5 loop). For instance, three deletion mutations $\Delta 69-70$, $\Delta 142-144$ and $\Delta 211$ are part of o variant, which are reported to escape most neutralizing mAbs.[30,31] Notably, not only the size of these four RDRs has expanded, but also new RDRs have appeared in SOI as shown in Fig. 2 and Fig. S5. In total, 5 more RDRs are found (3 in NTD, 1 in RBD and 1 in furin-cleavage sites Table S1). Mutation hotspots and a.a. sites with high $\boldsymbol{s}$ positively correlate with deletion occurrence in NTD Fig. S6. For example, a.a. sites of RDRs 69-70 and 144 exhibit the highest $\boldsymbol{s}$ value across full spike glycoprotein Fig. 2. Mutations in these regions are featured in VoC/VoI as presented in Table S1. Comparing genome sequences of δ variant to reference one, we deduce that R158G is substantially a consequence of a 6-nt deletion in RNA genome(deletion of $AGTTCA$ in the middle of $GAGTTCAGA$ leads to GGA, encoding Gly). In other words, deletion of neclueotides could also introduce 'substitution' of the neighbor a.a. site in protein, suggesting similar cases such as $\Delta 246-252$ D253N in λ variant and Δ142-143 Y145D in o variant.

The spike is found to be highly glycosylated and contains 22 N-linked glycans,[42] which is added to the spike glycoprotein druing a protein modification process called glycosylation and has crucial roles in viral pathobiology.[43] The heavily glycosylated surface of can interfere with antibody recognition as these glycans are able to act as a shield and mask antibody epitopes, resulting in immune escape.[44] As an introduction of glycosylation could make a successful shielding from mAbs, mutation to N could be very important. In deed, we found T20N (γ), Y145N (μ) and D253N (λ), together with the existing conserved glycan sites in NTD (bottom panel of Fig. 2), these mutations could result in higher escape rate from mAbs. After checking the locations of 7 NTD RDRs in 3D structure, we noticed that these regions are mostly bordered by four glycans N17, N74, N122 and N149, which is highly overlapped with the com-

mon surface that has been targeted by many antibodies.[37] Those regions are 14-20(N1 loop), 140-158(N3 loop) and 245-264(N5 loop), which corresponding to RDR 13-27, RDRs 136-151 & 156-158 and RDR 241-252, respectively.[36]

As seen in Fig. 2, RDR13 − 27 (N1 loop) exhibits high **s**, indicating its high mutation activity. Mutations of three VoC have been successfully captured as SOI: L18F(β, γ) T19R(δ) T20N(γ) P26S(γ). L18F is associated with a reduction in binding for mAb.[36] T20N could possibly introduce more complete glycosylation as discussed above. We noticed that region 14-20 are reported as one of the most common targeted sites, where mutations or deletions that alter epitopes conformation could help to increase antigenicity. Similarly, all the rest RDRs in NTD locate in highly-accessible regions[45], thus easily targeted by antibodies. Sites 21, 27 and 29 could be the next potential mutation region, as we see that this RDR is keeping expandingFig. 2.

$\Delta 69 - 70$, core of RDR67 − 77 (N2 loop), emerged in Aug. 2020 and massively spread as part of α variant and is reported to partially rescue infectivity of spike glycoproteins. Study has also found a $\Delta 69 - 70$ mutant which acquired N439K and Y453F escape mutations later lead to increased spike incorporation and mediate faster kinetics of cell-cell fusion than wild-type with only D614G.[46,47] In vitro, one spike double mutant bearing both $\Delta 69 - 70$ and D796H conferred modestly decreased sensitivity to convalescent plasma, while maintaining infectivity levels that were similar to the wild-type virus.[37,45,48] Mutation in either border of this region have been reported as part of VoC/VoI(A67V(η, o) G75V&T76I(λ)), which is reported to harm mAbs binding in o variant.[30] In our results, site 77 also shows high level of concern as it may result in similar effects. Study has also report mutation T95I, G142D and $\Delta 142 - 145$ reduce neutralizing antibodies targeting.[30] Sites 97, 98 and 112 are also included in our SOI as there is T95I(o, ι, μ) nearby and 112 has been showing a increasing trend in the spreading mutation cluster from the observation in Emerging Variants in GISAID.[9]

Region 140 − 158, located in N3 loop, has been classified as part of beta-hairpin of antigenic surpersite, a common shared epitopes targeted by many mAbs,[36] such as 4A8 (targeting 144-147 and 150). Studies has shown that mutations or deletions in this region, such as Δ141–144&Δ146, W152C and H146Y,[28,36] could facilitate evasion from mAbs. A recent work on o variant with mutation $\Delta 142 - 144 \& Y145D$ also confirmed the vulnerability this region.[30] In spite of the relative conserved site N149, which provide crucial a glycan shielding for the region, most of the sites here show very high mutation diversity, as we can see from the $\boldsymbol{s}$ in Fig. 2 lower panel. The SAP for certain site could be optimized with more SAP sampled. For example, in the vicinity of site 150, study has claimed many escape mutations in 148, 150 and 151,[49] where they all exhibit high $\boldsymbol{s}$ (Fig. 2). We can also see the emergence of a new RDR: $\Delta 156 - 157$ $R158G$ as part of δ variant, where experiments has proved F157A a mAb evasion mutation.[50] Computational modeling predicted that the deletion and insertion in loops N3 and N5 prevent binding of neutralizing antibodies.[51] Together, 141 and 146 are the outlined emerged SOI remaining to be monitored Fig. 2.

RDR 241 − 252, in N5 loop, is another frequently mAb-targeted region[37] as there is no glycan protection nearby. Study has report a 11-residue insertion between Y248 and L249 that brings a glycan, completely abolished neutralization.[51] This region actually consists of two deletion peaks: $\Delta 241 - 243$ (β), and $\Delta 246 - 252$ (λ). Interestingly, our data collection has shown a sign of fusion of two RDRs since there is only 2 sites in between. In addition, some deletions also occurred in region 256-265 as we can see in Fig. 2 lower panel. As the ongoing evolution continues, a expanded RDR might be established which ranges from sites 241-252 till 265. Particularly, site 262 has already been showing a high $\boldsymbol{L}$-index.

Spike RBD mutations. Our method deLemus has successfully revealed a number of RBD mutations found in current variants as leading mutations. Mutations at this domain are normally associated with altering the binding of virus to host cells and hence the transmissibility. Mutation N501Y, one of the shared sites in Fig. 3 as it is established in the α, β, γ, μ and o variants has been shown to enhance the binding affinity of the RBD to ACE2, as the mutation from asparagine to aromatic tyrosine is simulated to strengthen binding by allowing ring stacking interactions.[52] Apart from improved binding, the N501Y modification in the α variant is sufficient to reduce the activity of several mAbs.[53,54]

For another shared site 484, mutations at this site have been demonstrated in studies to not only reduce the neutralization titers of covalescent plasma by manifold but also lower the efficiency of monoclonal antibodies.[49,55,56] Mapping of mutations has revealed E484K as the strongest escape site for class 2 antibodies,[34] which largely represent those found in covalescent polyclonal plasma. Not only antibody evasion, structural studies using cryo-EM reveals that E484K in combination with N501Y and D614G, found in β, γ and o can enhance ACE2-RBD binding by inducing local rearrangements involving rotamer placement of H34 of ACE2 and Q493 of RBD.[57]

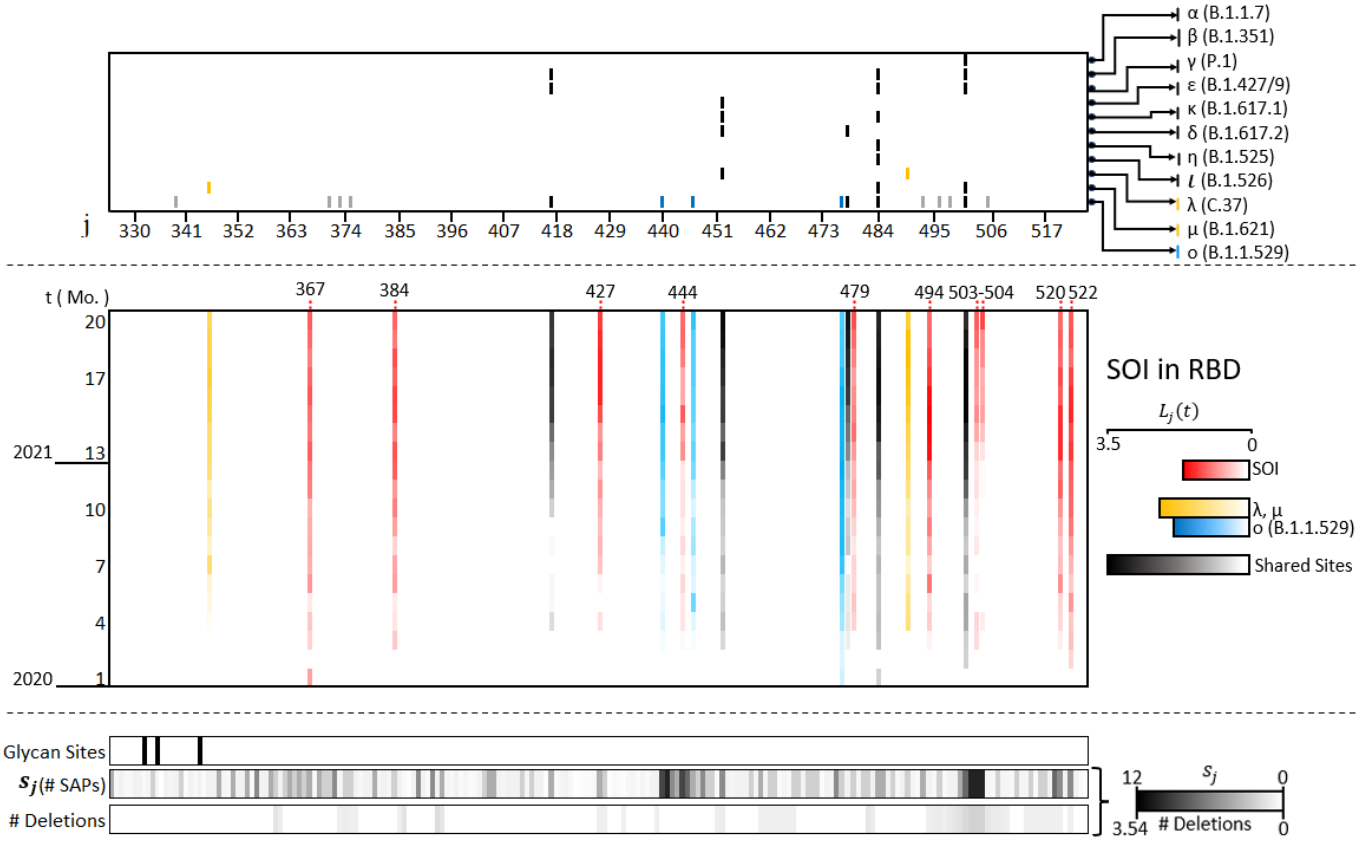


Fig. 3. Top panel: RBD mutations of the spike protein of VoC/VoI defined by WHO. Sites shared by more than one VoC/VoI are colored in black. Middle panel: Dynamic expedition of SOI in RBD. Outlined emergent mutated sites are are highlighted by a.a. site indices above. Bottom panel: Glycan sites, s_j (substitution number of a.a. at site j), and deletion occurrence are the mean values from the sequences from Mar. to Aug. 2021.

Mutation at another leading mutation site 417 has also been claimed to have significant disruptive effects to the binding of antibodies. It is the strongest escape mutant for class 1 antibodies[34] and is able to escape a potential treatment antibody.[34] K417N-E484K-N501Y ("triple mutant") mutations have been found to inhibit the binding of the powerful class 2 RBD mAbs DH1041 and DH1043.[58] Molecular Dynamics simulations showed that when coupled to ACE2, the "triple mutant" causes the most conformational changes in RBD compared to E484K or N501Y alone, supporting the convergent evolution and selection benefit of the alterations from this "triple mutant". Besides that, L452R found in the δ, κ, ϵ, and λ variants is an antigenically significant substitution which has been demonstrated to decrease neutralisation by a variety of monoclonal antibodies[54,59,60] and convalescent plasma.

A recently announced VoC, the o variant is reported to contain 15 RBD mutations, inclusive of almost all of the sites found in previous variants that are revealed by deLemus of which 417, 484 and 501 have been discussed above. Although the o variant was declared by WHO in late Nov,[61] those mutations have already showed up at an earlier time as leading mutations in our results. One of them is at site 478, which has emerged in δ variant previously and subsequently joined by site 477 to appear in o variant. Site 477 also exhibits high SAPs possibly due to the continuous evolution process that prompts for more SAPs. In fact, mutations at 477 (S477G, S477N, and S477R) have been identified as several mAb escape variants. Among them, mutant S477G is able to show resistance to two of the four sera examined.[60] Along with 477, residues 476, 478, 487, and 586 of the closed RBD clockwise next to the upright RBD are blocked by the upright RBD in the open spike structure.[45] Mutation at 477 may therefore impede the binding of certain antibodies that preferentially bind to the RBD in closed spike structure. In terms of binding, S477 shows the highest flexibility and MD simulations revealed that S477G and S477N increase the SARS-COV-2 spike's binding to the hACE2 receptor.[62]

Sites 440 and 446 found in o variant also shows a high **s** value, suggesting these sites are actively mutating during evolution. Perhaps triggered by antibodies, both of them are in the loop formed by residues 440-449 and combined together

they reduce the binding of antibodies.[31] Steric hindrance is observed for interactions with antibodies REGN10987 and 2-7.[30] For 446, this site appears at a highly antigenic region of the spike structure, and changes at this site have been shown to influence neutralisation by both mAbs and antibodies present in polyclonal serum.[34,59,60] Both mutations N440K and G446S hence confer strong antibody resistance to the o variant. Also, they located peripherally to the binding site of RBD to ACE2.[49] N440K is found to be one of the spike mutations that clusters around human ACE2 recognition sites showing high affinity for ACE2. These spike regions are structurally flexible, allowing for optimization of spatial binding arrangements.[63]

Realising the impacts of mutations in the RBD, we suggest that our approach should be promising in outlining SOI that is very likely to hold mutations that is going to have similar effects on viral fitness. Simulations have shown that RBD has to undergo a transition from a "down" to "up" state to be open for binding.[64] SOI D427, along with D405 and R408 play a key role in coordinating and sustaining the opening of glycan gate N343 which supports the RBD in its "up" conformation state. Mutation of N343 into other residues was shown to cause a considerably decrease in ACE2 binding. It is also found that glycosylation deletion of N343 significantly decreased infectivity, demonstrating the critical role of glycosylation in viral infection.[59] On the contrary, there is a possibility that residues can mutate to better coordinate spike opening and facilitate the binding to ACE2. Given the assisting role of 427 in the process of spike opening, we assume this leading mutation to have high possibility to emerge as a significant mutation.

Previous mutations have suggested the tendency for the spike glycoprotein to pick up RBD mutations associated with antibody escape. With a rising population acquiring immunity through vaccination or infection, under the selective pressure of immune responses, we propose that the spike's evolutionary pathway will continue to favour evasion of antibodies. For 444, this SOI contains high SAPs and is suggested to be subjected to antibody pressure considering its high structure-based antibody accessibility scores for both the closed and open RBD.[45] Also, this site should be given specific attention due to its proximity to mutation sites 440 and 446 in o variant. Besides that, mutation P384L, combined with mutations K417N, E484K and N501Y, is able to evade all vaccines because the highest antibody disruption count is observed.[65] Another SOI 494 has been shown to contain mutation S494P that is disruptive to the binding of antibodies and a potential candidate for vaccine escape.[65,66]

Interestingly, it is plausible that in addition to increasing transmissibility, mutations that enhance the receptor binding affinity independently without affecting antibody recognition of epitopes could be an indirect way of promoting immune escape.[45] The a.a. site 503 that is engaged in forming polar interactions with ACE2 receptor with improved interactions induced by mutation N501Y,[67] along with its adjacent site 504 could be promising candidates. Apart from that, both of them also exhibit high SAPs and we have seen how sites (440, 446) with high SAPs actually did emerge in o variant. The a.a. sites 520 and 522 give relatively stable signals as leading mutations and both of them have also been measured to have increased binding affinity, indicating favourable RBD to ACE2 binding.[65] Lastly, 367 as an active leading mutation up till recent months, might be deemed significant as the mutation V367F has been shown to be able to enhance expression and also the thermal stability of mammalian-expressed RBD leading to better binding towards ACE2.[68]

Conclusion

Our method deLemus made it possible to monitor such a complex and dynamical system, protein sequence samplings, in a systematic way by extracting crucial mutation information from massive time-dependent data that takes into account the amplitude and space of single site mutation. Most of the VoC/VoI mutations has been located as SOI from enormous data sets. The outlined potential mutation sites are reported to serve as a guideline in deciphering the extremely complicated mutational pathways. The rapid evolution of the spike protein is evident from the increasing trend observed for both the number of mutated a.a. sites and also the number of SAPs over time. As the most mutations in SOI are already in VoC/VoI, adequate attention must be given to our novel outlined SOI. Besides, our results have comprehended previously characterised RDRs[28] and see newly evolved RDRs with forewarning from of $\boldsymbol{L}$-index, highlighting the importance of deletion in SARS-CoV-2 evolution. Collectively, our results show a high occurrence of SAPs and deletions in NTD, especially in common mAbs targeted region.[36,37] The disproportionately high rate of mutation in NTD implies an evolutionary route in which the virus retains optimal ACE2 binding effectiveness while retaining the flexibility necessary for immune escape. For RBD, we observe that mutations from previous VoC/VoI being incorporated into the new o variant, probably because these mutations are linked to fitness enhancing traits. Given the crucial role of RBD, which already shows robust binding affinity with interacting with ACE2, we propose that the emergence of SOI are driven by immunological response. As a result, mutations in RBD are likely to increase the virus's transmissibility and susceptibility to antibodies, affecting existing pandemic control measures.[30]

Understanding the ongoing evolution of SARS-COV-2 is essential to control the ongoing global pandemic. Therefore, collective and sustained global efforts have to be called for to ensure a more comprehensive system of surveillance. All of these are essential to achieve positive developments in understanding the evolutionary progress of the virus and lead to more effective management of the pandemic.

Acknowledgements

We are grateful to Professor Yong Huang for many thought-provoking discussions and encouragement in the course of this work, which is supported in part by Hong Kong Branch of South Marine Science and Engineering Guangdong Laboratory (SMSEGL20SC01) (for H.S.), HKUST grant (R9418) and Society of Interdisciplinary Research (SOIRÉE). This research made use of the computing resources of the X-GPU cluster supported by the Hong Kong Research Grant Council Collaborative Research Fund (C6021-19EF). We would also like to acknowledge the enlightening advice from Jie Gao in the early stage of this work; the technical support from Long T. Ta, Siqi Lei, and Shihong Chen for improving figures; Leon Ritchie, Ruichen Li, and Salina Lam in data collection and preparation.

References

1. Zhou, P., Yang, X.-L., Wang, X.-G., *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579,** 270–273 (2020).
2. *WHO Coronavirus (COVID-19) Dashboard.https://covid19.who.int/*
3. Woo, P. C. Y., Huang, Y., Lau, S. K. P., *et al.* Coronavirus Genomics and Bioinformatics Analysis. *Viruses* **2,** 1804–1820 (2010).
4. Dearlove, B., Lewitus, E., Bai, H., *et al.* A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc. Natl. Acad. Sci. USA.* **117,** 23652–23662 (2020).
5. Denison, M. R., Graham, R. L., Donaldson, E. F., *et al.* Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol.* **8,** 270–279 (2011).
6. Minskaia, E., Hertzig, T., Gorbalenya, A. E., *et al.* Discovery of an RNA virus 3→5 exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proc. Natl. Acad. Sci. USA.* **103,** 5108 (2006).
7. Vilar, S. & Isom, D. G. One Year of SARS-CoV-2: How Much Has the Virus Changed? *Biology* **10,** 91 (2021).
8. Goldstein, S. A., Brown, J., Pedersen, B. S., *et al.* Extensive recombination-driven coronavirus diversification expands the pool of potential pandemic pathogens. *bioRxiv. https://doi.org/10.1101/2021.02.03.429646* (2021).
9. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22,** 30494 (2017).
10. Rambaut, A., Holmes, E. C., O'Toole, Á., *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5,** 1403–1407 (2020).
11. *Tracking SARS-CoV-2 variants.https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/*
12. Berenbaum, M. R. On COVID-19, cognitive bias, and open access. *Proc. Natl. Acad. Sci. USA.* **118** (2021).
13. Su, H., Zhou, F., Huang, Z., *et al.* Molecular Insights into Small-Molecule Drug Discovery for SARS-CoV-2. *Angew. Chem. Int. Ed.* **60,** 9789–9802 (2021).
14. Benson, D. A., Cavanaugh, M., Clark, K., *et al.* GenBank. *Nucleic Acids Res.* **41,** D36–D42 (2012).
15. Benton, D. J., Wrobel, A. G., Xu, P., *et al.* Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* **588,** 327–330 (2020).
16. Renhong, Y., Yuanyuan, Z., Yaning, L., *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367,** 1444–1448 (2020).
17. Walls, A. C., Tortorici, M. A., Frenz, B., *et al.* Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. *Nat. Struct. Mol. Biol.* **23,** 899–905 (2016).
18. Walls, A. C., Park, Y.-J., Tortorici, M. A., *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181,** 281–292.e6 (2020).
19. Hoffmann, M., Kleine-Weber, H. & Pöhlmann, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Mol. Cell* **78,** 779–784.e5 (2020).
20. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5,** 562–569 (2020).
21. Lan, J., Ge, J., Yu, J., *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581,** 215–220 (2020).
22. Chen, R. E., Zhang, X., Case, J. B., *et al.* Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. *Nat. Med.* **27,** 717–726 (2021).
23. McCormick, K. D., Jacobs, J. L. & Mellors, J. W. The emerging plasticity of SARS-CoV-2. *Science* **371,** 1306–1308 (2021).
24. Wu, F., Zhao, S., Yu, B., *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579,** 265–269 (2020).
25. Goujon, M., McWilliam, H., Li, W., *et al.* A new bioinformatics analysis tools framework at EMBL–EBI. *Nucleic Acids Res.* **38,** W695–W699 (2010).
26. Golub, G. H. & Van Loan, C. F. *Matrix Computations* 4th ed. (The Johns Hopkins University Press, Maryland, 2013).
27. Ng, P. C. & Henikoff, S. Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annu. Rev. Genomics. Hum. Genet.* **7,** 61–80 (2006).
28. McCarthy, K. R., Rennick, L. J., Nambulli, S., *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* **371,** 1139–1142 (2021).
29. Motozono, C., Toyoda, M., Zahradnik, J., *et al.* SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. *Cell Host Microbe* **29,** 1124–1136.e11 (2021).
30. Liu, L., Iketani, S., Guo, Y., *et al.* Striking Antibody Evasion Manifested by the Omicron Variant of SARS-CoV-2. *Nature. https://doi.org/10.1038/d41586-021-03826-3* (2021).
31. Cao, Y. R., Wang, J., Jian, F., *et al.* B. 1.1. 529 escapes the majority of SARS-CoV-2 neutralizing antibodies of diverse epitopes. *Nature. https://doi.org/10.1038/d41586-021-03796-6* (2021).
32. Cameroni, E., Bowen, J. E., Rosen, L. E., *et al.* 1 Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron 2 antigenic shift. *Nature. https://doi.org/10.1038/d41586-021-03825-4* (2021).
33. Greaney, A. J., Starr, T. N., Barnes, C. O., *et al.* Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat. Commun.* **12,** 4196 (2021).
34. Starr, T. N., Greaney, A. J., Addetia, A., *et al.* Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **371,** 850–854 (2021).

35. Callaway, E. *et al.* Beyond Omicron: what's next for COVID's viral evolution. *Nature* **600,** 204–207 (2021).
36. McCallum, M., De Marco, A., Lempp, F. A., *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184,** 2332–2347.e16 (2021).
37. Cerutti, G., Guo, Y., Zhou, T., *et al.* Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. *Cell Host Microbe* **29,** 819–833.e7 (2021).
38. Chi, X., Yan, R., Zhang, J., *et al.* A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* **369,** 650–655 (2020).
39. Wang, N., Sun, Y., Feng, R., *et al.* Structure-based development of human antibody cocktails against SARS-CoV-2. *Cell Res.* **31,** 101–103 (2020).
40. Voss, W. N., Hou, Y. J., Johnson, N. V., *et al.* Prevalent, protective, and convergent IgG recognition of SARS-CoV-2 non-RBD spike epitopes. *Science* **372,** 1108–1112 (2021).
41. Puray-Chavez, M., LaPak, K. M., Schrank, T. P., *et al.* Systematic analysis of SARS-CoV-2 infection of an ACE2-negative human airway cell. *Cell Rep.* **36,** 109364 (2021).
42. Yasunori, W., Allen Joel, D., Daniel, W., *et al.* Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **369,** 330–333 (2020).
43. Watanabe, Y., Bowden, T. A., Wilson, I. A., *et al.* Exploitation of glycosylation in enveloped virus pathobiology. *Biochim. Biophys. Acta. Gen. Subj.* **1863,** 1480–1497 (2019).
44. Zhao, X., Chen, H. & Wang, H. Glycans of SARS-CoV-2 Spike Protein in Virus Infection and Antibody Production. *Front. Mol. Biosci.* **8,** 53 (2021).
45. Harvey, W. T., Carabelli, A. M., Jackson, B., *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19,** 409–424 (2021).
46. Meng, B., Kemp, S. A., Papa, G., *et al.* Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Rep.* **35,** 109292 (2021).
47. Telenti, A., Arvin, A., Corey, L., *et al.* After the pandemic: perspectives on the future trajectory of COVID-19. *Nature* **596,** 495–504 (2021).
48. Kemp, S. A., Collier, D. A., *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592,** 277–282 (2021).
49. Weisblum, Y., Schmidt, F., Zhang, F., *et al.* Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife* **9** (2020).
50. Suryadevara, N., Shrihari, S., Gilchuk, P., *et al.* Neutralizing and protective human monoclonal antibodies recognizing the N-terminal domain of the SARS-CoV-2 spike protein. *Cell* **184,** 2316–2331.e15 (2021).
51. Andreano, E., Piccini, G., Licastro, D., *et al.* SARS-CoV-2 escape from a highly neutralizing COVID-19 convalescent plasma. *Proc. Natl. Acad. Sci. USA.* **118,** e2103154118 (2021).
52. Dejnirattisai, W., Zhou, D., Supasa, P., *et al.* Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell* **184,** 2939–2954.e9 (2021).
53. Wibmer, C. K., Ayres, F., Hermanus, T., *et al.* SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat. Med.* **27,** 622–625 (2021).
54. Wang, Z., Schmidt, F., Weisblum, Y., *et al.* mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature* **592,** 616–622 (2021).
55. Greaney, A. J., Loes, A. N., Crawford, K. H. D., *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29,** 463–476.e6 (2021).
56. Baum, A., Fulton, B. O., Wloga, E., *et al.* Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* **369,** 1014–1018 (2020).
57. Mannar, D., Saville, J. W., Zhu, X., *et al.* Structural analysis of receptor binding domain mutations in SARS-CoV-2 variants of concern that modulate ACE2 and antibody binding. *Cell Rep.,* 110156 (2021).
58. Saunders, K. O., Lee, E., Parks, R., *et al.* Neutralizing antibody vaccine for pandemic and pre-emergent coronaviruses. *Nature* **594,** 553–559 (2021).
59. Li, Q., Wu, J., Nie, J., *et al.* The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **182,** 1284–1294.e9 (2020).
60. Liu, Z., VanBlargan, L. A., Bloyet, L.-M., *et al.* Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host Microbe* **29,** 477–488.e4 (2021).
61. *Update on Omicron.https://www.who.int/news/item/28-11-2021-update-on-omicron*
62. Singh, A., Steinkellner, G., Köchl, K., *et al.* Serine 477 plays a crucial role in the interaction of the SARS-CoV-2 spike protein with the human receptor ACE2. *Sci. Rep.* **11,** 4320 (2021).
63. Gan, H. H., Twaddle, A., Marchand, B., *et al.* Structural Modeling of the SARS-CoV-2 Spike/Human ACE2 Complex Interface can Identify High-Affinity Variants Associated with Increased Transmissibility. *J. Mol. Biol* **433,** 167051 (2021).
64. Sztain, T., Ahn, S.-H., Bogetti, A. T., *et al.* A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nat. Chem.* **13,** 963–968 (2021).
65. Wang, R., Chen, J., Hozumi, Y., *et al.* Emerging vaccine-breakthrough SARS-CoV-2 variants. *ArXiv,* arXiv:2109.04509v1 (2021).
66. Wang, R., Chen, J., Gao, K., *et al.* Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries. *Genomics* **113,** 2158–2170 (2021).
67. Santos, J. C. & Passos, G. A. The high infectivity of SARS-CoV-2 B.1.1.7 is associated with increased interaction force between Spike-ACE2 caused by the viral N501Y mutation. *bioRxiv,* 2020.12.29.424708 (2021).

68. Starr, T. N., Greaney, A. J., Hilton, S. K., *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182,** 1295–1310.e20 (2020).

Supplementary Information

Methodology of Dynamic Expedition of Leading Mutations (deLemus)

With the multiple sequence alignment data we collected, we have constructed mutation matrix $\boldsymbol{H}(t)$ from each month from Jan. 2020 to Sep. 2021, where t represents the tth month from Jan. 2020. For a month that have m new individual sequence, the system is represented by a $m \times n$ matrix $\boldsymbol{H}(t)$. Different rows of the matrix represent sites mutation information in different sequence. Then, the sequence correlation matrix writes:

$$\boldsymbol{C}_{seq}(t) = \boldsymbol{H}^T(t) \cdot \boldsymbol{H}(t) \tag{1}$$

where $\boldsymbol{C}_{seq}(t)$ is an $n \times n$ matrix with n eigenvectors. The matrix composed by eigenvectors of sequence correlation matrix writes:

$$\boldsymbol{N}^T(t) = [\boldsymbol{N}_1^T(t), \boldsymbol{N}_2^T(t), \ldots, \boldsymbol{N}_n^T(t)] \tag{2}$$

Moreover, $\boldsymbol{H}(t)$ can be factorized according to SVD[1] as:

$$\boldsymbol{H}(t) = \boldsymbol{P}(t) \cdot \boldsymbol{\Sigma}(t) \cdot \boldsymbol{N}(t) \tag{3}$$

where $\boldsymbol{P}(t)$ and $\boldsymbol{N}(t)$ share the same eigenvalue series $\boldsymbol{\Sigma}(t)$. The contribution of the ith eigenvector of $\boldsymbol{N}(t)$ to system is $\sigma_i^2(t)/\sum_i \sigma_i^2(t)$. Then we define site mutation series $\boldsymbol{N}_{ij}(t)$ for each month as:

$$\boldsymbol{N}_{ij}(t) = (\boldsymbol{\Sigma}(t) \cdot \boldsymbol{N}(t))_{ij} \tag{4}$$

$\boldsymbol{N}_{ij}(t)$ represent the mutation amplitude of jth site corresponding to the ith eigenvector. After we apply SVD on these matrix, we have calculated the eigenvalue series σ_i and mutation series $\boldsymbol{N}_{ij}(t)$ of SARS-CoV-2 shows as below. Here we only keep the top 4 $\boldsymbol{N}_{ij}$ series. For different mutation set i, we define the $\boldsymbol{L}_{ij}$ index for each mutation set as the product of number of SAPs and mutation amplitude:

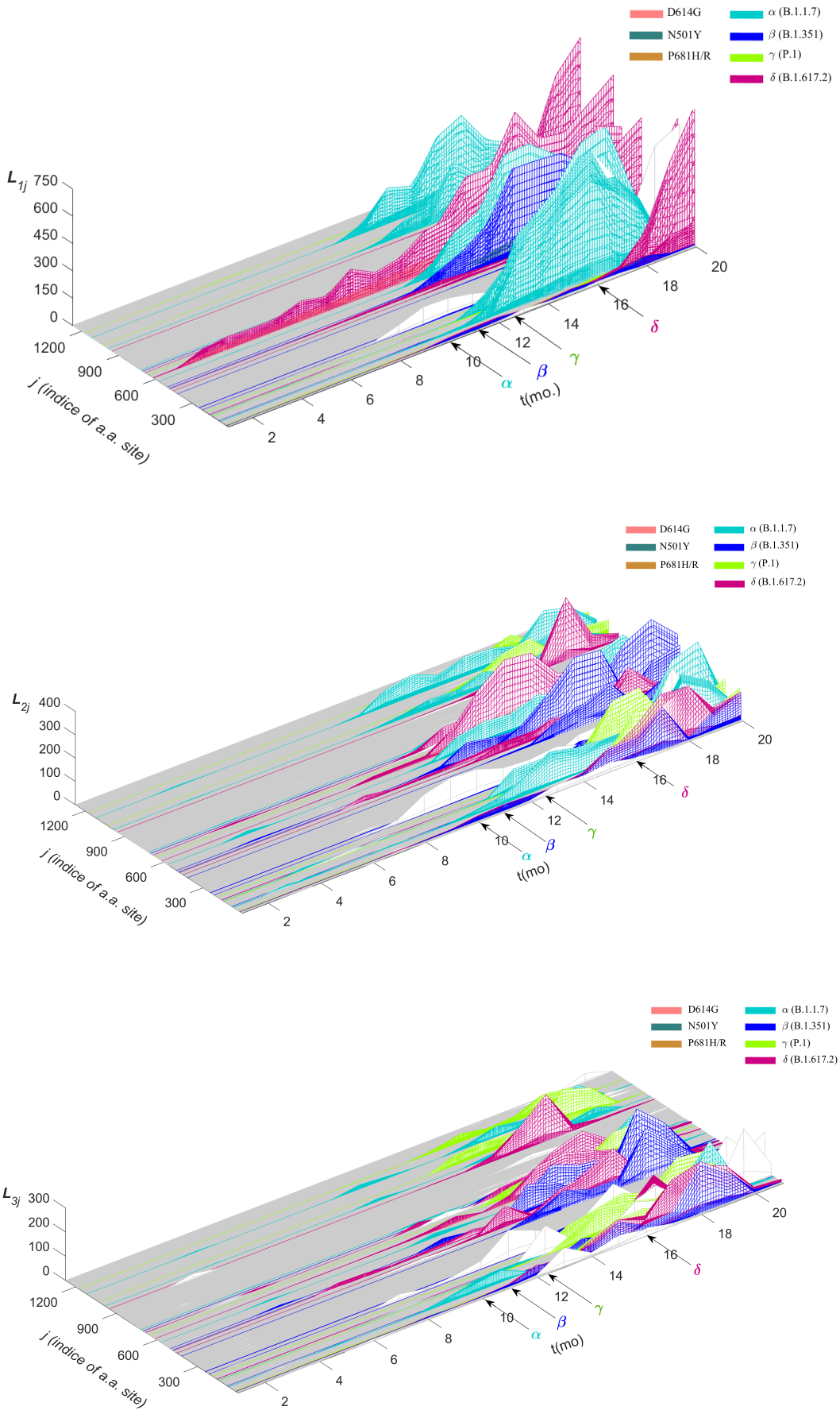
$$\boldsymbol{L}_{ij}(t) = \boldsymbol{s}_j(t) \cdot \boldsymbol{N}_{ij}^2(t) \tag{5}$$

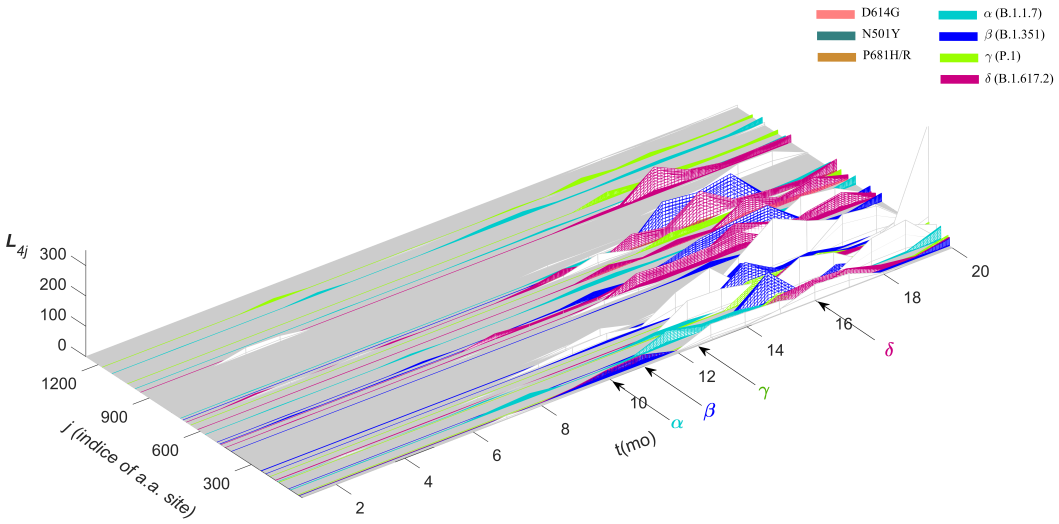
The time evolution of the leading 4 $\boldsymbol{L}_{ij}$ index is showed in Fig. S1(a), where different variants are labeled by different colors. The contribution of $\boldsymbol{L}_{ij}(t)$ to the mutation index of SARS-CoV-2 is $\sigma_i^2(t)/\sum_i \sigma_i^2(t)$. To get the overall effect of mutation of SARS-CoV-2, we define the $\boldsymbol{L}$-index for each jth site as the magnitude of the summed index of the top 4 mutation sets. In other words, the $\boldsymbol{L}$-index of the jth site for the tth month is define as:

$$\boldsymbol{L}_j(t) = \sqrt{\sum_i^4 \boldsymbol{L}_{ij}^2(t)} = \boldsymbol{s}_j(t) \cdot \sqrt{\sum_{i=1}^4 \boldsymbol{N}_{ij}^2(t)} \tag{6}$$

Dynamic expeditions of SOI are presented in Fig. S1(a). Composition of top 6 leading mutation sets is showed at Fig. S1(b), where we can see the contribution of 5th ans 6th mutations sets are already negligible. Thus, only top 4 mutation sets are considered in this work.

(a)


D614G
N501Y
P681H/R
α (B.1.1.7)
β (B.1.351)
γ (P.1)
δ (B.1.617.2)
L1j
750
600
450
300
150
0
1200
900
600
300
j (indice of a.a. site)
2
4
6
8
10
12
14
16
18
20
t(mo.)
α
β
γ
δ
L2j
400
300
200
100
0
t(mo)
L3j



(b)

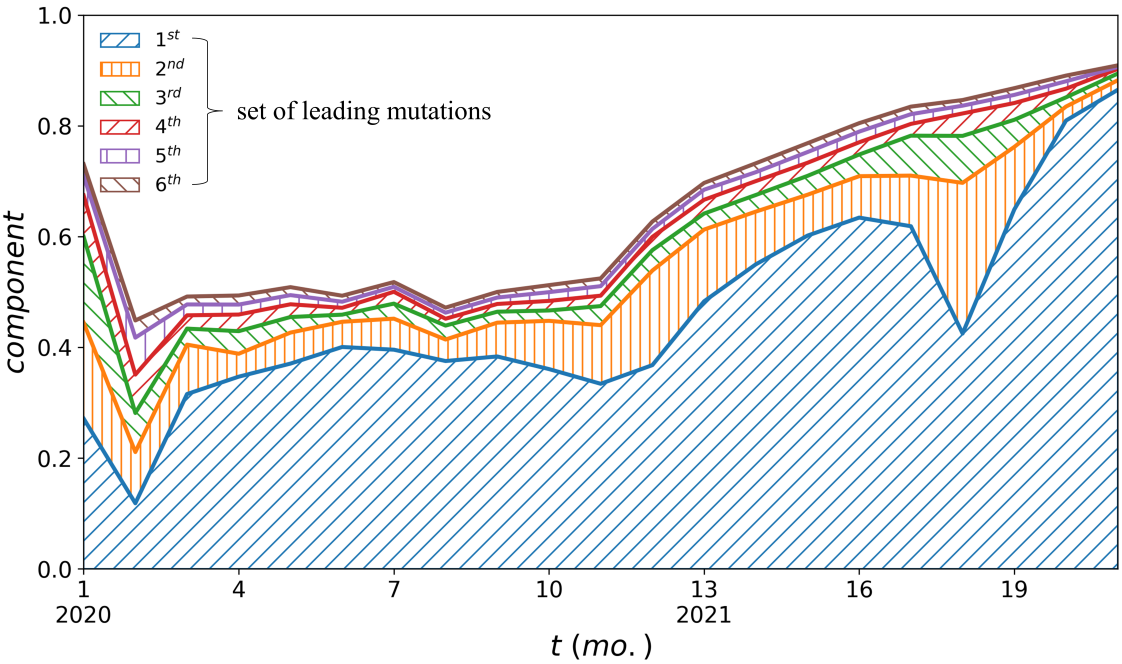
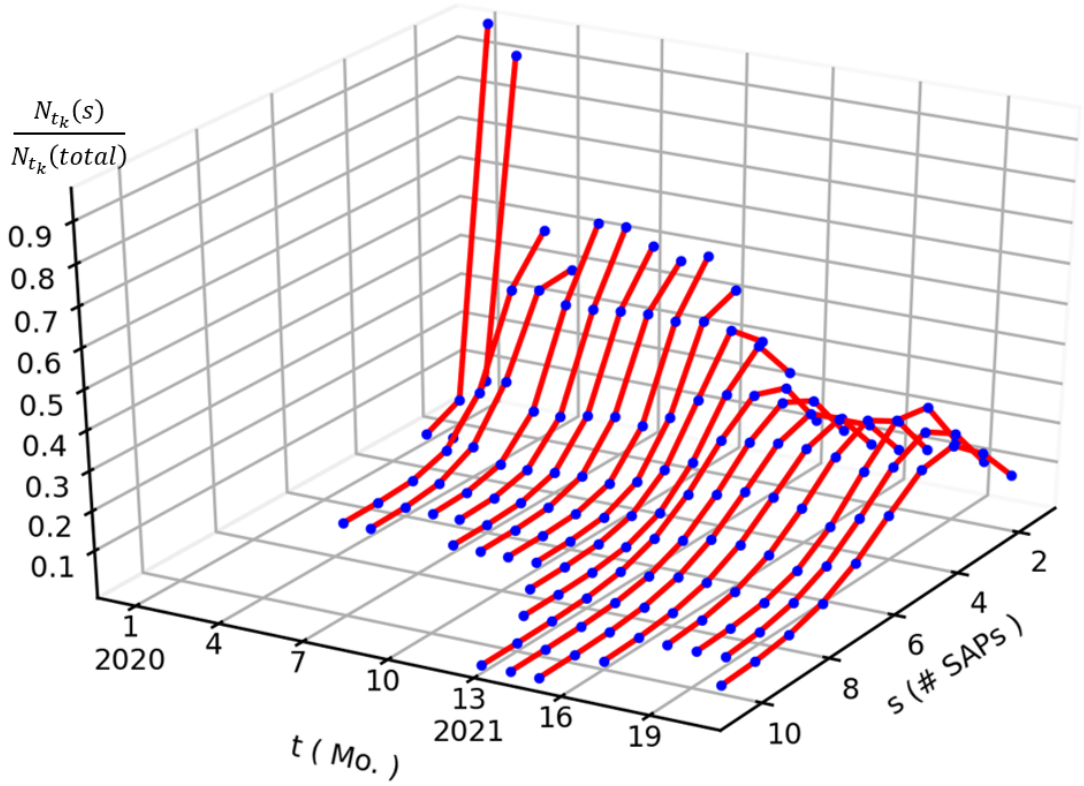


Fig. S1. (a). Time evolution top 4 mutation series. Shortly after the outbreak of the epidemic, 614 have quickly mutated and becomes the most common mutation from Jan. to Oct. in 2020. At the early Oct. 2020, sites that related to α variance becomes more active and shows a potential for new variance. After that, α strain emerges and becomes the most widely spread SARS-CoV-2 variant at the winter of 2020. Although β and γ were reported at Nov. and Dec. 2020 respectively, these two strains are suppressed by α variant and didn't show up till the second mutation set. At Apr. 2021, sites related to δ strain shows up and become the dominant strain in the first mutation set soon after it was reported, indicating the appearance of a new variant. Moreover, sites of VoC also show up but with lower mutation amplitude score. (b). Composition of top 6 leading mutation sets

(a)



(b)

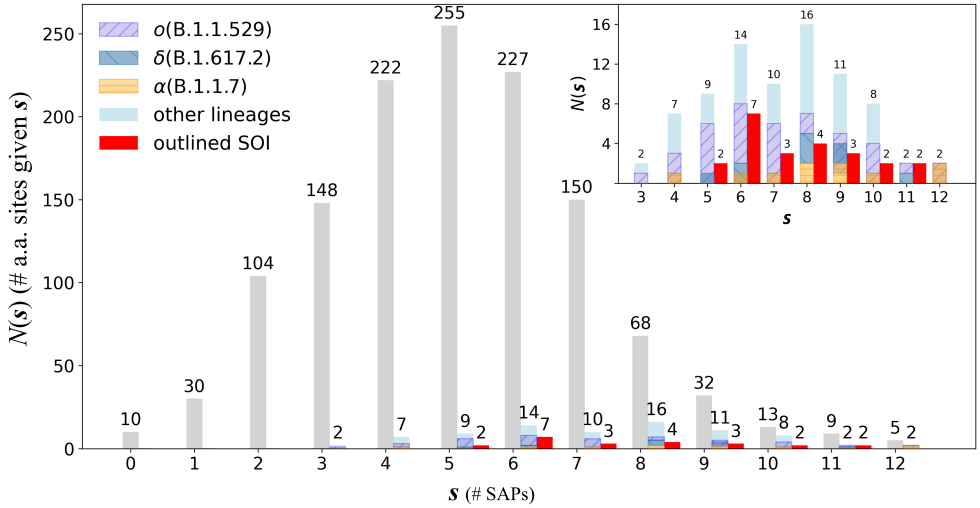


Fig. S2. (a). The dynamic of $N(\boldsymbol{s})$ distribution over the months. (b). Overall $N(\boldsymbol{s})$ distribution inside the time window from Jan. 2020 to Aug. 2021. The $N(s)$ is plotted in grey color, VoC/VoI is showed in blue. The composition of different strain is showed at top right corner of the plot with different color: purple for o, dark blue for δ and yellow for α. The outlined SOI are showed in red. (b). Overall $N(\boldsymbol{s})$ distribution. Both α, δ and o show a preference of the region with more SAP options. most VoC/VoI sites of SARS-CoV-2 loacte in the right side of the mean value and nearly half of the sites have been reported as VoC/VoI exhibit a **s** value large than 9. Moreover, as showed in Fig. S6, sites with high **s** are more likely to harbour deletion mutations and play crucial roles in the evolution of SARS-CoV-2.

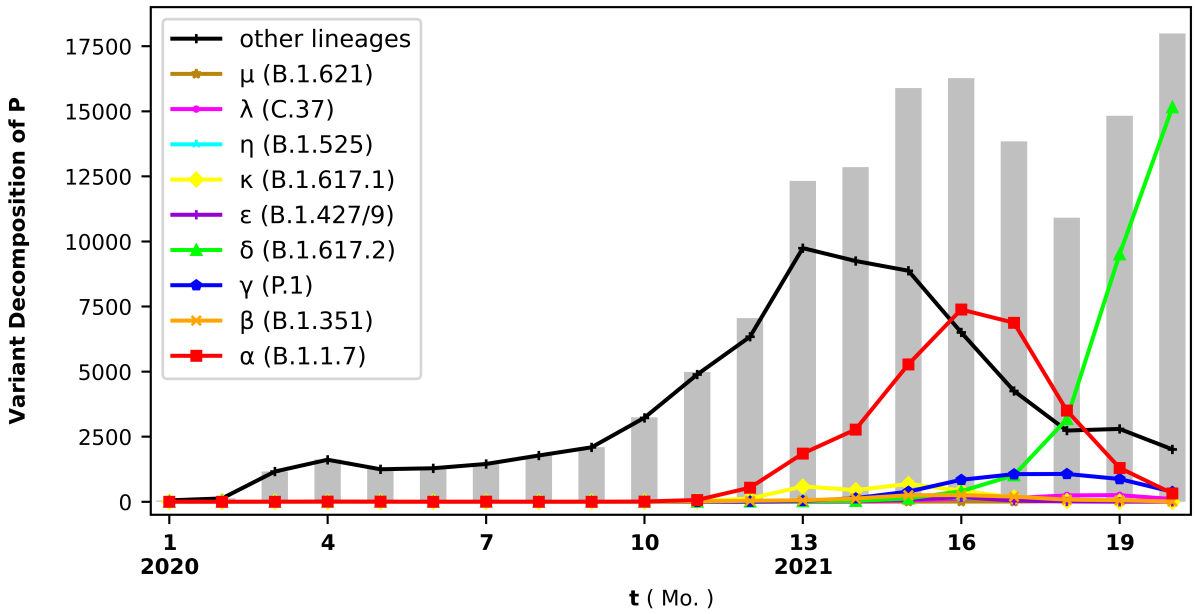


Fig. S3. Distribution for non-degenerate sequences are plotted in grey. Decomposition of all sequences is shown in color lines. α and δ variant are the dominant lineages. The composition of different lineages is indicative of the diversity of SARS-CoV-2 at certain time point, similar trend has been observed in Fig. 1(b).

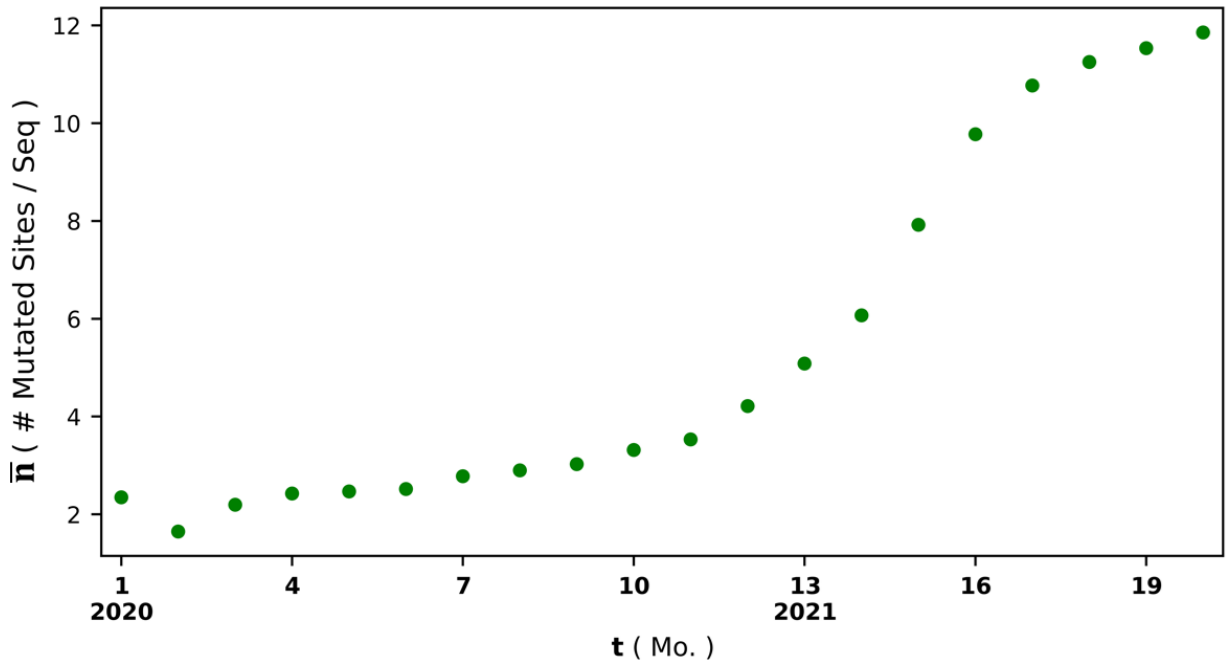


Fig. S4. $\bar{\boldsymbol{n}}_{t_k} = [\sum_{ij} \boldsymbol{H}_{ij}/m(t_k)$, where $H_{i,j}(t_k)$ is the mutation matrix and $m(t_k)$ is number of sequence in the kth month. $\bar{\boldsymbol{n}}$ reflects the ensemble average deviation of all sequence compared to the initial sequence, which suggest the increments of mutated a.a. sites per month is around $0.5\tilde{2}$ (Fig. 1(a)). As the curve shows a acceleration around Jan. to Apr. 2021, which is indicative of that new evolution stage, the dominance of α and δ variant.

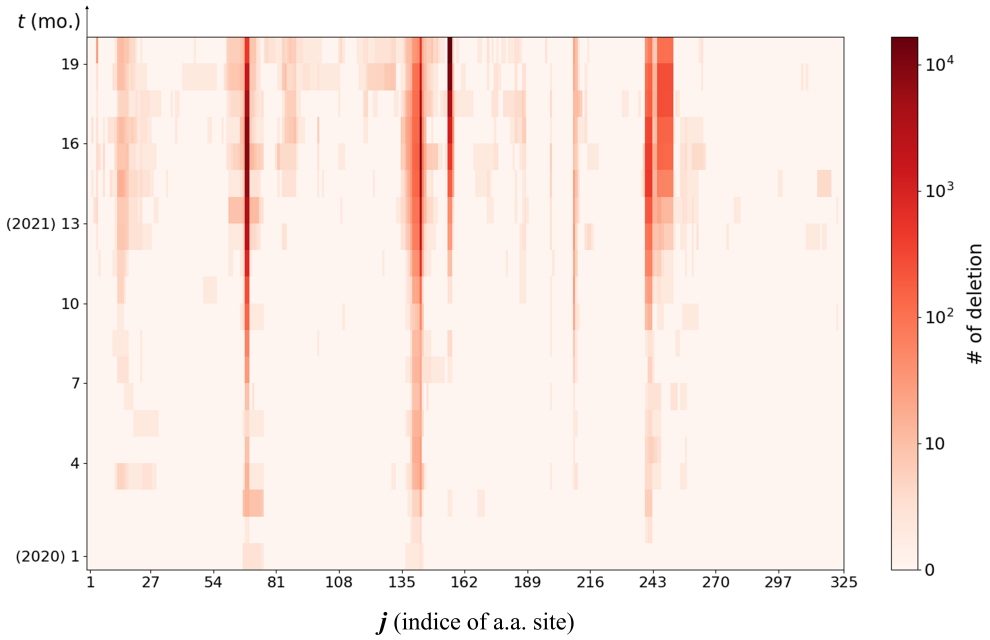


Fig. S5. Evolution of deletion regions in NTD. Among the four previously identified RDRs: $\Delta 69-70$ (N2 loop), $\Delta 141-144$ (N3 loop), $\Delta 210$ (between N4 and N5 loop), and $\Delta 243-244$ (N5 loop)[3]). Not only these RDRs are getting broadened but also new RDRs are emerging Fig. S5. In total, 3 more RDRs are found in NTD, 1 more in RBD and 1 more in furin-cleavage sites (Table S1)

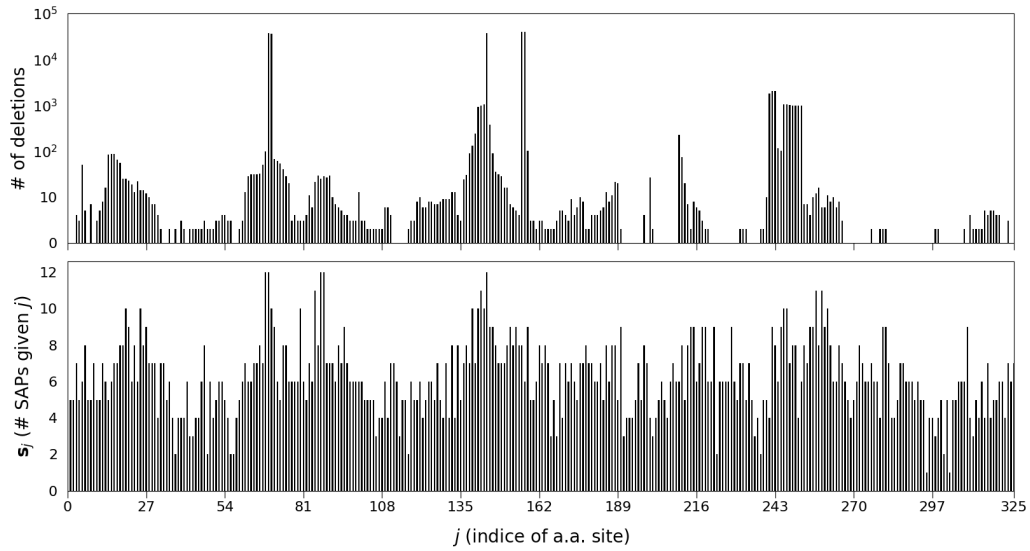


Fig. S6. Top: Distribution of deletion occurrences in NTD; Bottom: Distribution of $\mathbf{s}$ (Number of SAPs) distribution. As we can see their peaks locate at similar positions, which indicates a positive correlation between deletions and SAPs, suggesting the evolutionary pressure in that particular region.

RDRs	Related Mutations in VOC/VOI
13-27	L18F(β,γ) T19R(δ) **T20N**(γ) P26S(γ) Δ24-26&A27S (BA.2)
61-76	A67V(η,ο) Δ69-70(α,ο,BA.2) G75V&T76I:λ
83-91	
136-151	D138Y(γ) Δ142-143&Y145D(ο) Δ144(α,ο,BA.2,η) Y144S&**Y145N**(μ)
156-158	Δ156-157 and R158G
210-212	Δ211&L212I(ο)
241-252	Δ241-243(β) Δ246-252&**D253N**(λ) D253G(ι)
500-504	N501Y (α,β,γ,ο,BA.2,μ) Y505H(ο,BA.2)
676-680	N679K(ο,BA.2) P681H(a,ο,BA.2) P681R(δ,κ)

Shared epitopes
14-20
140-158
245-264

Glycan sites
N17, N61, N74, N122, N149, N165, N234, N282, N331, N334, N343

TABLE S1: Left panel: RDRs and related mutations in VoC/VoI. The first 4 RDRs matches well with previous work, with some extended regions[3]. The time evolution of RDRs can been seen in Fig. S5; Right panel: Top: Shared epitopes regions as reported on the previous studies[4][5], which is also highly overlapped with RDRs. Bottom: Glycan sites in NTD. the shared epitopes regions sre broderred by four glycan sites N17, N74, N122 and N149.[5][6] T20N, Y145N and D253N are bold as such mutations may introduce glycosylations.

Confirmed SOI			
NTD			RBD
5:ι	95:ι,μ,ο	215:β	346:μ
18:β,γ	138:γ	242:β	417:β,γ,ο
19:δ	142:ο	243:β	440:ο
20:γ	143:ο	246:λ	446:ο
26:γ	144:α,η,μ,ο	247:λ	452:δ,κ,ε,λ
67:η,ο	145:μ,ο	248:λ	477:ο
69:α,η,ο	152:ε	249:λ	478:δ,ο
70:α,η,ο	156:δ	250:λ	484:β,γ,κ,η,ι,μ,ο
75:λ	157:δ	251:λ	490:λ
76:λ	158:δ	252:λ	501:α,β,γ,μ,ο
80:β	190:γ	253:ι,λ	

Outlined SOI	
NTD	RBD
21	367
27	384
29	427
77	444
97	479
98	483
112	494
141	503
146	504
153	520
262	522

TABLE S2: Left: Confirmed SOI, corresponding variants are labeled in Greek letter for each site. Almost all of mutation sites from VoC/VoI in the NTD are identified by $\boldsymbol{L}$-index, only four are missing: 13,52,154 and 241. For RBD, all mutation sites from VoC/VoI are identified by $\boldsymbol{L}$-index. Right: Outlined SOI, possible mutational sites to be tested in the future variants. Overall, there are 11 sites for each NTD and RBD reported in this work, respectively.

References

1. Golub, G.H. & Van Loan, C. F. Matrix Computations 4th ed., (The Johns Hopkins University Press, Maryland, 2013).
2. Walls, A. C., Park, Y.-J., Tortorici, M. A., et al. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glyco-protein. Cell 181, 281–292.e6 (2020).

3. McCarthy, K. R., Rennick, L. J., Nambulli, S., et al. Recurrent deletions in the SARS-CoV-2 spike glycoprotein driveantibody escape. Science 371, 1139–1142 (2021).
4. Cerutti, G., Guo, Y., Zhou, T., et al. Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domaintarget a single supersite. en. Cell Host Microbe 29, 819–833.e7 (2021).
5. McCallum, M., De Marco, A., Lempp, F. A., et al. N-terminal domain antigenic mapping reveals a site of vulnerabilityfor SARS-CoV-2. en. Cell 184, 2332–2347.e16 (2021).