bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

Muhammad Hasan*,[1,2] Zhouyi He*,[1,2] Mengqi Jia*,[1] Alvin C. F. Leung*,[1,3] Kathiresan Natarajan*,[4] Wentao Xu,[1] Shanqi Yap,[1] Feng Zhou,[1] Shihong Chen,[1] Hailei Su†,[5] 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]*Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou, China*

[3]*Division of Life Science, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China*

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

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

(Dated: April 6, 2023)

Throughout the coronavirus disease 2019 (COVID-19) pandemic, the continuous genomic evolution of its etiological agent, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has generated many new variants with enhanced transmissibility and immune escape capabilities. Being an essential mediator of infections and a key target of antibodies, mutations of its spike glycoprotein play a vital role in modulating its evolutionary trajectory. Here, we present a time-resolved statistical method, Dynamic Expedition of Leading Mutations (deLemus), to analyze the evolutionary dynamics of the SARS-CoV-2 spike. Together with analysis of its single amino acid polymorphism (SAP), we propose the use of $\boldsymbol{L}$-index in quantifying the mutation strength of each amino acid site, such that the evolutionary mutation pattern of the spike glycoprotein can be unravelled.

I. Introduction

Ever since the emergence of coronavirus disease 2019 (COVID-19) in December 2019,[1] its rampant propagation has greatly hampered global socioeconomic activities, all the while leaving billions infected and millions dead.[2] Tremendous efforts have been put into mitigating the effects of this devastating pandemic by the international community, like implementing social lockdowns and mass vaccination programs. Regrettably, these measures have been ineffective in curtailing the persistence of this disease. The constantly shifting epidemiology of COVID-19 since its initial outbreak has been the result of the continuous evolution of its etiological agent, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Despite possessing one of the largest genomes among RNA viruses,[3] coronaviruses exhibit comparatively lower substitution rates than their viral counterparts, as they are capable of encoding a 3′-to-5′ exoribonuclease located at their nonstructural protein 14 for RNA-proofreading,[4–6] thereby facilitating proper genome replication by interacting with the low-fidelity RNA-dependent RNA polymerase.[7,8] However, this proofreading mechanism is not flawless, as the exoribonuclease lacks the ability to repair nucleotide deletions, which translationally alters the amino acid sequences of viral proteins.[9] Aside from the accumulation of deletions, another driving force by which coronaviruses evolve is RNA recombination.[10,11] By admixing the genomic content of genotypically dissimilar coronaviruses, it is possible for novel phenotypes to be introduced into the viral recombinants.[10,12] These evolutionary mechanisms have therefore enabled SARS-CoV-2 to amass a wide range of mutations as the pandemic progresses.

The immense selection pressures exerted upon SARS-CoV-2 have prompted its rapid genetic diversification, from which over 2500 lineages have been generated.[13] Within the first two years of the COVID-19 pandemic, the World Health Organization (WHO) had already designated five SARS-CoV-2 variants as variants of concern, namely alpha (α, B.1.1.7), beta (β, B.1.351), gamma (γ, P.1), delta (δ, B.1.617.2), and omicron (o, B.1.1.529), as well as numerous other variants of interest.[14] In rapid succession, each of these reported variants (RVs) emerged, proliferated, and outcompeted its antecedent, resulting in wavelike resurgences of COVID-19 cases. The emergence of all these variants

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

has introduced many novel mutations that continue to enhance the fitness of the virus.[15] Mutations of this nature, while uncommon,[16] have substantially complicated the development of therapeutics, which greatly hinders the progress of COVID-19 treatment research.[17] Hence, in order to develop more effective treatments and disease control strategies, it is of paramount importance to understand how this virus evolves over time.

Across the mutational landscape of the vast SARS-CoV-2 genome, the spike glycoprotein gene sits atop its plateau. The spike glycoprotein is a trimeric type I viral fusion protein that binds SARS-CoV-2 virion particle to the angiotensin converting enzyme 2 (ACE2) receptor on a host cell.[18] In terms of its structure, the spike glycoprotein is composed of 2 subunits, the N-terminal subunit 1 (S1) and C-terminal subunit 2 (S2), within which multiple domains lie. The S1 region mainly facilitates ACE2 binding,[19,20] and is made up of an N-terminal domain (NTD), a receptor-binding domain (RBD), and 2 C-terminal subdomains (SD1 and SD2), while the downstream S2 region is responsible for mediating virus-host cell membrane fusion.[21] Being a key mediator of cellular infections via both cell surface and endosomal entry pathways,[18] the spike glycoprotein is a primary target for antibodies,[22,23] immune effector cells,[24,25] and COVID-19 therapeutics.[26–29] Due to these selection pressures, the spike glycoprotein exhibits much higher mutational activities than other SARS-CoV-2 structural proteins, whose mutations are capable of altering its infectivity and antigenicity.[17,30] With its functional significance, elucidating the evolutionary trajectory of the spike glycoprotein would be the first step in deciphering the mechanisms behind virus evolution as a whole.

Nonetheless, even with extensive international collaborations in clinical and laboratory investigations, research on understanding SARS-CoV-2 spike glycoprotein evolution is still in its infancy. The complete sequence space of the spike glycoprotein, formed by all possible amino acid combinations of 1273 residues, encompasses more than 20^{1000} different sequences. However, its most recent evolutionary pathway envelops only a very small fraction of the entire sequence space, comprising a few million ($\sim 10^6$) distinct sequences only. While data retrieved from Global Initiative on Sharing Avian Influenza Data (GISAID) reveals that nearly all spike glycoprotein residues have mutated at least once since the initial COVID-19 outbreak, only few of these amino acid sites exhibit polymorphism. In other words, most identified spike glycoprotein sequences are concentrated within a small region of the whole sequence space, and that majority of the possible sequences remain unexplored. This shows that, even though the evolutionary process of the spike glycoprotein is highly dynamic, particular mutational constraints exist to prevent the spike glycoprotein from navigating through all the possible sequence options. In this work, a method named Dynamic Expedition of Leading Mutations (deLemus) is developed to quantitatively characterize the robust properties of the SARS-CoV-2 spike glycoprotein evolutionary dynamics, from which variations in both mutation rate and single amino acid polymorphism (SAP) can be identified. With each domain of the spike glycoprotein possessing divergent functions that are crucial in facilitating the infection process of SARS-CoV-2,[27] $\boldsymbol{L}$-index from our proposed deLemus analysis has been employed to quantify the mutational activity of each amino acid site, where it has been demonstrated that $\boldsymbol{L}$-index is capable of effectively outlining novel leading mutations (LMs) before their corresponding emerging variants are designated as RVs by the WHO.

II. Methodology

Detecting dynamic patterns from big data sets has always been a major challenge in data analysis. In this work, we propose one method, dubbed as deLemus, to investigate the evolutionary dynamics of the SARS-CoV-2 spike glycoprotein at an amino acid sequence level.

A collection of 7,940,305 SARS-CoV-2 spike glycoprotein amino acid sequences was downloaded from the GISAID hCoV-19 database on February 9th 2022.[31] We used EPI_ISL_402124 as the reference sequence of the spike glycoprotein.[1,32] As there exists a substantial amount of repeated sequences in the original data, all degenerate sequences have been removed to keep only the non-degenerate ones for further analysis of sequence mutations. Overall, 172,280 non-degenerate sequences have been retrieved from the total set of reported sequences uploaded to GISAID between December 2019 and December 2021.

Sequences submitted within the same month were grouped together. Multiple sequence alignment was then consecutively conducted on each group using Clustal Omega to check the occurrence of substitution or deletion at each amino acid site,[33] with respect to the reference sequence. This yielded the total number of mutated sites in all sequences $\boldsymbol{n}$ and the number of sequences $P(\boldsymbol{n})$ with a given $\boldsymbol{n}$ in each month. The mutation rate Ξ in the unit of seq^{-1}mo^{-1} was calculated based on the

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

total number of mutations per sequence per month (Fig. S2). The total number of SAP at each jth amino acid site in each tth month $\boldsymbol{s}_j(t)$ and the number of amino acid sites $N(\boldsymbol{s})$ were also computed, from which Poisson distribution was observed, giving the monthly average SAP number $\bar{\boldsymbol{s}}$ (Fig. S3).

For each tth month, one $m \times n$ mutation matrix $\boldsymbol{H}(t)$ is constructed based on the multiple sequence alignment data, where m is the number of non-degenerate sequences displayed in a particular month, and n is the length of the spike glycoprotein amino acid sequence. In other words, each row represents one non-degenerate sequence from that month, and each column corresponds to one residue in the sequence. For the ith sequence, if the jth residue is changed, the corresponding matrix component $\boldsymbol{H}_{ij}(t)$ would be set to 1. Otherwise, it would be set to 0.

We then factorized $\boldsymbol{H}(t)$ by singular value decomposition,[34]

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

where $\boldsymbol{P}(t)$ is an $m \times m$ matrix and $\boldsymbol{N}(t)$ is an $n \times n$ matrix. (see Supplemental Material for details).

From the monthly $\boldsymbol{\Sigma}_i(t)$ and $\boldsymbol{N}_{ij}(t)$, we collected the top 4 leading sets of mutations to calculate the $\boldsymbol{L}$-index $\boldsymbol{L}_j(t)$, which is used to quantify the mutation strength of each jth site in each tth month,

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

The amino acid sites were arranged according to their $\boldsymbol{L}_j(t)$, and top-ranked ones were identified as LMs of each month from January 2020 to December 2021.

III. Results and Discussions

The structural information of each confirmed LM site was presented via AlphaFold2,[35] as shown in Fig. 1. Here, we highlighted the mutation sites that have appeared in RVs. Based on the different characteristics of each protein domain, we divided the outlined LMs into four regions for discussion: NTD, RBD, SD1 and SD2, and S2. With the structural information provided in Fig. 1, functional changes of the spike glycoprotein conferred by these LMs can be illustrated.

A. N-Terminal Domain (NTD)

The NTD is an S1 ectodomain situated at the outermost region of the SARS-CoV-2 spike glycoprotein, on which several epitopes lie.[16,36–38] While the NTD does not directly interact with ACE2 receptors, the domain's close spatial proximity to the RBD has enabled some of its mutations to alter the cell entry dynamics of SARS-CoV-2.[39,40] With their abilities to vastly impact the antigenicity and infectivity of the virus, it is of utmost importance to closely monitor the evolutionary trajectory of the NTD (Fig. 2).

Domains across the SARS-CoV-2 spike glycoprotein generally adopts substitution as a common mode for generating mutations, which would be discussed thoroughly in later parts. However, particularly in the NTD, insertions and deletions are also prevalent mechanisms by which mutations are generated, leading to its structural plasticity.[16,41] Deletions in this domain have been characterized to frequently occupy particular locations of the spike gene known as RDRs, where partial nucleotide removals within specific stretches of codons can give rise to new nucleotide arrangements.[9,42] One such example would be a persistent six-nucleotide deletion ΔAGTTCA within the spike gene segment GAGTTCAGA that results in the formation of the mutated codon CGA, replacing $_{156}$EFR$_{158}$ with a single glycine residue (R158G) in the δ variant. Interestingly, with regards to NTD deletions, data from our deLemus analysis shows that lengths of these RDRs expand over time (Fig. 2 and Fig. S5).

The overwhelming mutational activity of this domain can be clearly demonstrated by how the emergence of each RV introduces novel mutable sites in the NTD. Our investigation began with the $\boldsymbol{L}$-index calculation of each NTD amino acid site starting from December 2019 to outline potential LMs. By November 2020 (11th mo.), we identified six LMs among numerous candidates, all of which were later confirmed to be present in the α and β variants that emerged in December 2020. These included all three mutations in the α NTD (ΔH69/V70 and ΔY144) and three mutations in the β NTD (L18F, D80A, and D215G), as presented in Fig. 1 and Fig. 2. The three α mutations are situated within previously documented RDRs (RDR1 and RDR2),[9] where the former double deletion has been reported to promote infectivity by enhancing syncytia formation,[43] while the latter single deletion located

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

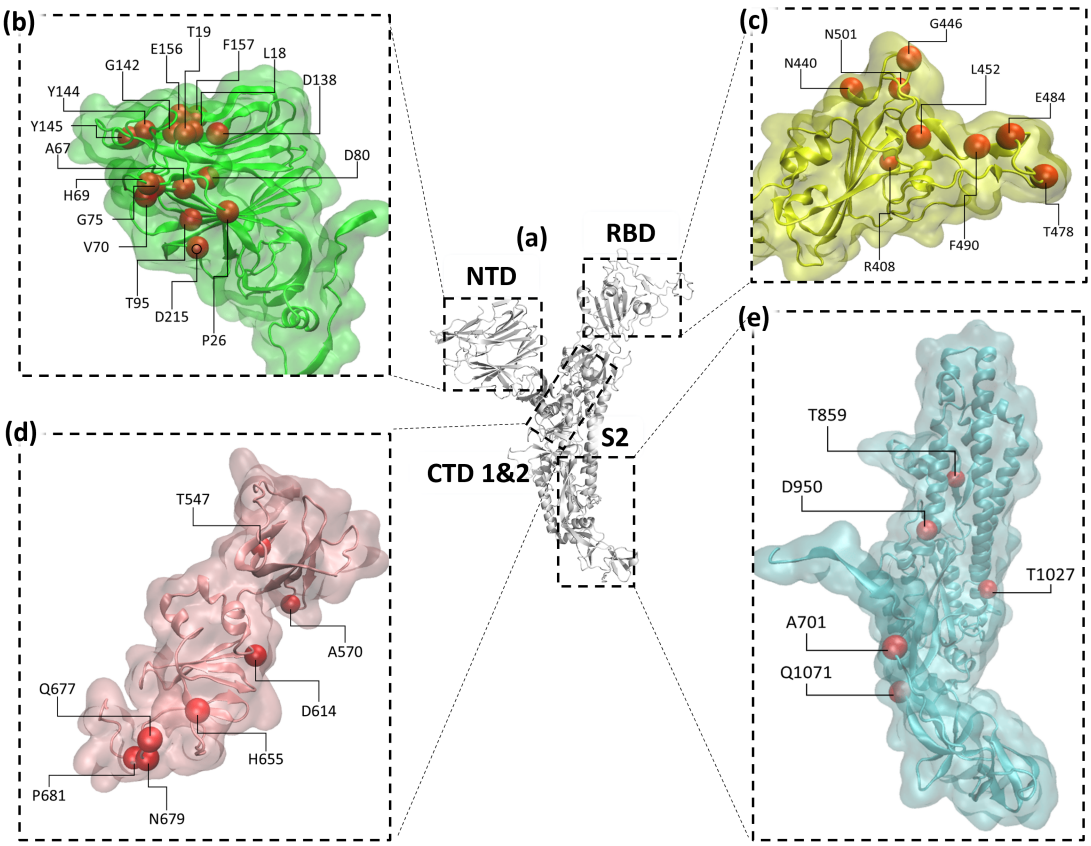

Fig. 1. Confirmed LM sites (red spheres) in different domains of the SARS-CoV-2 spike glycoprotein. (a) A SARS-CoV-2 spike glycoprotein monomer (generated by AlphaFold2, Colab version, using the reference sequence EPI_ISL_402124) encompassing the four major functional domains: NTD, RBD, SD1 and SD2, and S2. (b) Mutation sites outlined in NTD (green), most of which are clustered in the outer regions. (c) Mutation sites outlined in RBD (yellow), nearly all the captured mutations are located in the receptor-binding motif (RBM, 438-506). (d) Mutation sites outlined in SD1 and SD2 (pink). (e) Mutation sites outlined in S2 (cyan).

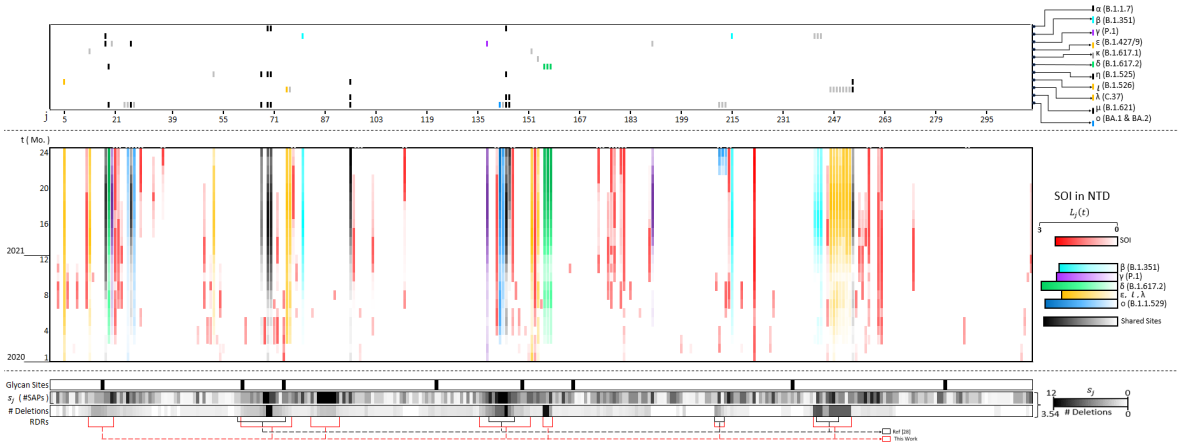

Fig. 2. Evolutionary trajectory of the NTD between January 2020 (1st mo.) and December 2021 (24th mo.). Top panel: NTD mutations harbored by RVs as defined by the WHO. Sites shared by more than 1 RV are colored in black. Middle panel: NTD LMs outlined by our deLemus analysis, whose mutational activities are measured by their $\boldsymbol{L}_j(t)$. Bottom panel: Distributions of glycosylation sites, $\boldsymbol{s}_j(t)$, and deletions in the NTD, as computed by their mean values of occurrence in all non-degenerate sequences collected between March to August 2021. Bracketed stretches indicate the locations of recurrent deletion regions (RDRs), where those reported in ref[9] are in black, and those identified in this work are in red.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

at the NTD N3 loop has been found to impair the neutralizing activities of several antibodies.[44,45] As for the recorded β NTD mutations, all three of them have been found to confer immune escape capabilities.[45,46]

In December 2020 (12th mo.), we identified multiple potential LMs in the NTD, two of which were later confirmed to be present in the γ variant that emerged in January 2021: P26S and D138Y. In the following month (13th mo.), we outlined additional LMs, two of which, L5F and T95I, were then reported in the iota variant (ι, B.1.526) that emerged in February 2021 (14th mo.). By the subsequent month (14th mo.), we identified several other LMs, one of which was later confirmed to be present in the eta variant (η, B.1.527) that emerged in February 2021. While the functions of many of these mutations are not well-understood, it has been reported that the two LMs present in the γ variant disrupt the epitope of monoclonal antibody (mAb) 159, leading to a significant reduction in its neutralizing activity.[47]

Similarly, we identified multiple LMs in March 2021 (15th mo.), three of which were later confirmed to be present in the δ variant that emerged in April 2021[48]: T19R and the ΔE156/F157 double deletion. Moreover, we identified a novel LM, G75V, which was later harbored by the lambda variant (λ, C.37) emerged in the same month as the δ variant. Interestingly, while the ΔE156/F157 double deletion has been experimentally shown to improve viral fitness by conferring immune evasion capabilities against NTD-targeting antibodies,[49] the G75V mutation has been shown to significantly decrease the infectivity of SARS-CoV-2.[50]

We identified more novel LMs in October 2021 (22nd mo.), two of which have been confirmed as novel NTD mutations carried by the first o sublineage, the BA.1 strain: ΔG142 and Y145D (Fig. 1 and Fig. 2). This variant first originated from South Africa in late November 2021.[51,52] While the effects regarding mutations at G142 are not well-studied, its location within the NTD antigenic supersite implicates that G142 mutations may alter spike-antibody interactions,[45] as one study has demonstrated that the G142D mutation is capable of conferring significant resistance against NTD-targeting mAbs.[53] As for Y145D, this mutation has been reported to reduce the neutralization potencies of convalescent sera.[54]

B. Receptor-Binding Domain (RBD)

The RBD is a S1 domain that not only plays an essential role in ACE2 recognition,[55,56] but also acts as a region of immunodominance targeted by around 90% of all plasma or serum neutralizing antibodies.[16,57,58] Mutations in this domain therefore commonly possess the abilities to alter virus-ACE2 or virus-antibody binding affinities,[59] which enable the generation of variants with higher transmissibility or immune escape capabilities.[16,60] In fact, as shown in Fig. 1, most mutations in this domain are located in the RBM that serves as a spike-ACE2 binding interface. These mutations would therefore undoubtedly affect the infectivity of the virus. With such significant functional implications, it is necessary to track RBD mutations in a temporal manner.

Since the emergence of the first two RVs, α and β, different SARS-CoV-2 variants have continued to acquire mutations in the RBD, including the latest o family. Several of these RBD mutations have been successfully captured by our deLemus analysis. One such important mutation is N501Y, which was first outlined as an LM (9th mo.) by our deLemus analysis, confirmed in the α variant, and subsequently retained in the β, γ, mu (μ, B.1.621.1), and o variants, as depicted in Fig. 3. Studies have shown that the N501Y mutation can enhance ACE2 binding affinity by introducing π-π interactions between N501Y of RBD and Y41 of ACE2.[47] In addition, this mutation has been found to decrease the neutralizing activities of some mAbs.[46,60]

E484K is another mutation that was first outlined as an LM (8th mo.) by our deLemus analysis, and later confirmed in the β variant. Mutations in this site have been demonstrated to not only significantly reduce the neutralization titers of convalescent plasma, but also lower the activities of mAbs.[61–63] Mapping of mutations has revealed that E484K is the strongest escape site for class 2 antibodies,[64] which largely represent those found in convalescent polyclonal plasma. Although some studies suggested that mutations of the E484 residue would lead to diminished electrostatic complementarity between the RBM and the ACE2 receptor,[65] many structural biology studies have shown that, when introduced with other mutations like the aforementioned N501Y, the E484K mutation can impart increased RBD-ACE2 binding. For instance, in the β, γ and o variants, the E484K-N501Y-D614G triple mutation is found to enhance RBD-ACE2 binding by inducing local rearrangements involving rotamer placements between Q493 of RBD and H34 of ACE2.[66]

Two other novel RBD mutations, L452R and T478K, were also outlined as LMs (11th mo. and 13th mo.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

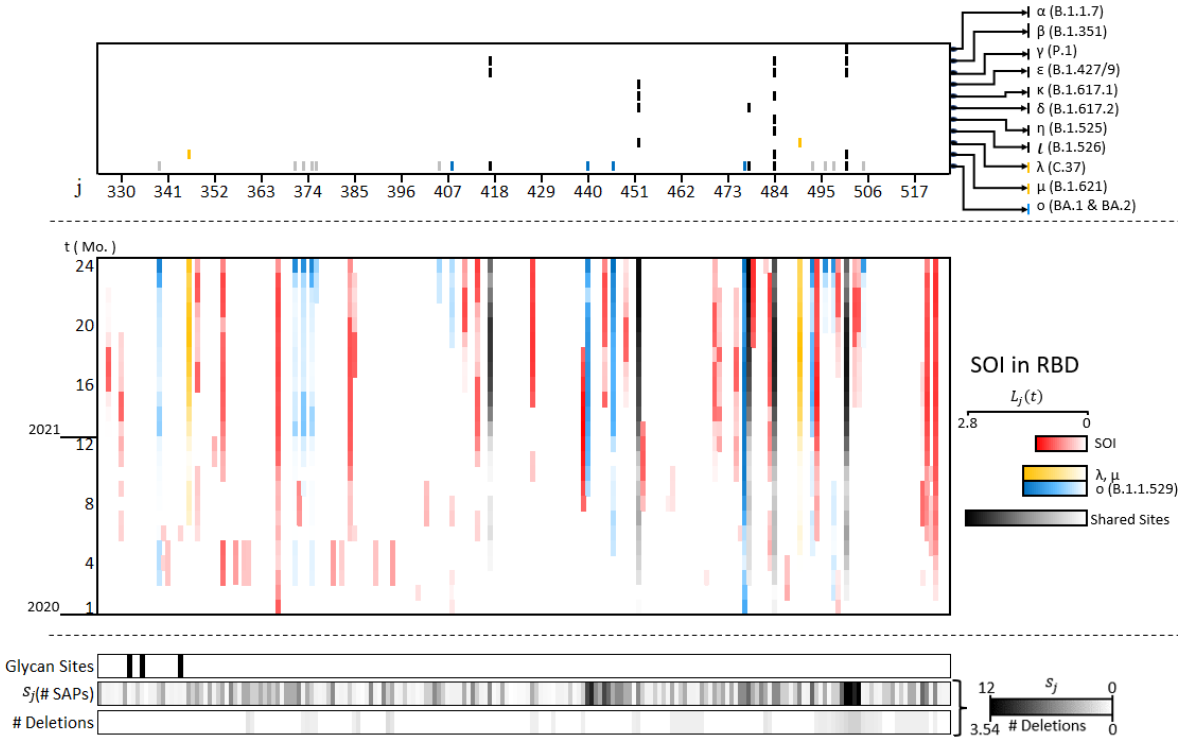

Fig. 3. Evolutionary trajectory of the RBD between January 2020 (1st mo.) and December 2021 (24th mo.). Top panel: RBD mutations harbored by RVs as defined by the WHO. Sites shared by more than 1 RV are colored in black. Middle panel: RBD LMs outlined by our deLemus analysis, whose mutational activities are measured by their $\boldsymbol{L}_j(t)$. Bottom panel: Distributions of glycosylation sites, $\boldsymbol{s}_j(t)$, and deletions in the RBD, as computed by their mean values of occurrence in all non-degenerate sequences collected between March to August 2021.

respectively) by our deLemus analysis. These mutations are later confirmed as fingerprint mutations of the δ variant that emerged in April 2021. Their locations within the epitope of several important neutralizing antibodies enable them to enhance the immune escape capabilities of the virus.[60,67–69] Furthermore, these two mutations have been reported to improve RBD-ACE2 interactions.[70–72] For the former mutation, computational studies have shown that variants possessing the L452R-E484Q-N501Y triple mutation exhibit a secondary structure rearrangement that is associated with an increase in RBD-ACE2 binding affinity.[73] As for the latter mutation, structural analysis has revealed that the T478K substitution allows the formation of two new hydrogen bonds located between Y489 of RBD and Y83 of ACE2, and F490 of RBD and K31 of ACE2 respectively, which results in tighter binding between the RBD and the ACE2 receptor.[72]

Known for its exceptionally high transmissibility, the *o* variant harbors a significant number of RBD mutations. Most of them have been reported in previous variants, but several new sites were outlined as LMs (22nd mo.) by our deLemus analysis, which encompasses the mutations at R408, N440 and G446. The R408S mutation has been shown to alter the antigenic property of the spike glycoprotein by disrupting the binding of F2 antibodies.[74] Unlike R408S, the N440K mutation can enhance spike-ACE2 binding affinity by increasing the electrostatic complementarity between the structurally flexible ACE2 recognition site and the ACE2 receptor.[75] As for the G446 residue which is situated at a highly antigenic region of the spike structure (Fig. 1), changes at this site have been shown to influence neutralization by both mAbs and antibodies present in polyclonal serum.[64,68,69]

C. Subdomains (SD1 & SD2) and S2 Subunit

The post-RBD region of the SARS-CoV-2 spike glycoprotein is consisted of SD1, SD2, and S2. While these regions do not directly engage the ACE2 receptor, they confer significant functions in spike allostery and post-ACE2-binding events respectively.[76,77] For the two SDs, their close spatial proximities to the NTD-RBD linker motif enable them to modulate RBD motion, where mutations in these regions could alter RBD "up-down" equilibrium.[78–80] S2, meanwhile, plays an important role in mediating host-virus membrane fusion by undergoing massive post-fusion conformational changes.[77,81] Hence, mutations in S2 can drastically impact the fusogenicity of the SARS-CoV-2 spike, both in

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

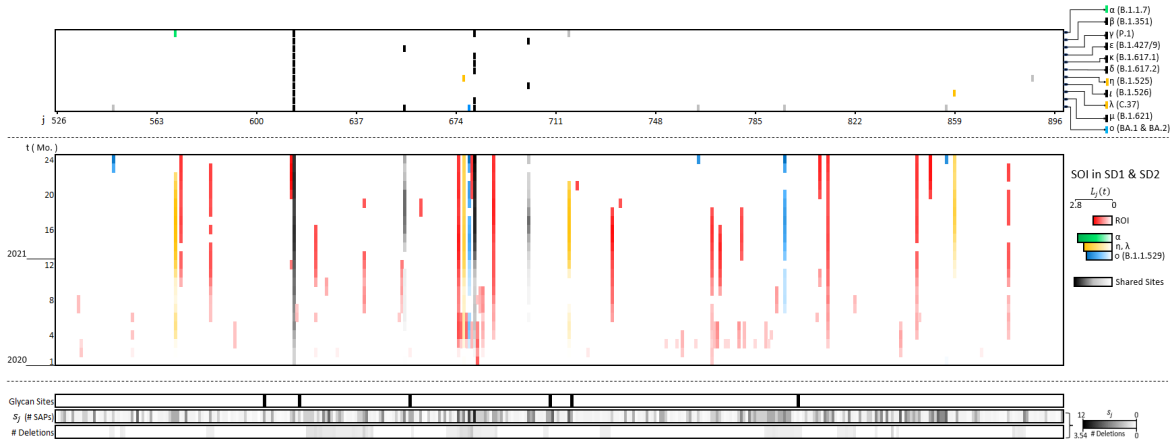

Fig. 4. Evolutionary trajectory of the SDs between January 2020 (1st mo.) and December 2021 (24th mo.). Top panel: SD mutations harbored by RVs as defined by the WHO. Sites shared by more than 1 RV are colored in black. Middle panel: SD LMs outlined by our deLemus analysis, whose mutational activities are measured by their $\boldsymbol{L}_j(t)$. Bottom panel: Distributions of glycosylation sites, $\boldsymbol{s}_j(t)$, and deletions in the SDs, as computed by their mean values of occurrence in all non-degenerate sequences collected between March to August 2021.

terms of host-virus membrane fusion for cell surface entry and cell-cell membrane fusion for syncytia formation.[18,26] Additionally, the post-RBD region houses the S1/S2 and S2' cleavage sites, which are proteolytically processed in order to facilitate membrane fusion processes.[21,81–83] The former site in particular is unique to SARS-CoV-2 within the subgenus *sarbecovirus*, as it is generated from a polybasic insertion at the C-terminal of SD2 ($_{681}$PRRAR↓S$_{686}$), and has been experimentally found to be an important determinant of viral transmission.[21,84] It is therefore quintessential to employ the deLemus method in tracking post-RBD mutations.

Like the previous sections, we first computed the $\boldsymbol{L}$-index of each post-RBD amino acid site in order to screen for potential LMs within SD1, SD2 and S2. By January 2020 (1st mo.), we had already outlined D614G as one of the LMs, a prominent single amino acid substitution proximal to the SD1-SD2 junction which has emerged from as early as April 2020 (Fig. 1 and Fig. 4).[85] Mechanistically, this mutation has been found to disrupt the interprotomer hydrogen bond involving S2 residues, thereby increasing the propensity for RBD to be in the "up" state.[86] Experimentally, it has been shown that the D614G mutation enhances S1/S2 cleavage, cell entry, replicative fitness, and transmissibility.[86–88] The increased viral fitness has therefore enabled D614G to be conserved in all RVs thereafter.

In October 2020 (10th mo.), we identified two LMs, A570D and P681H, both of which were later confirmed to be present in the α variant that emerged in December 2020. As shown in Fig. 1, these mutations are situated in the SD1 and SD2 region. In particular, the P681H mutation is located at the S1/S2 furin cleavage site and has been predicted to increase S1/S2 cleavability by substituting an electrically neutral proline with a cationic histidine, thereby enhancing the electrostatic complementarity between the cleavage motif and the anionic furin catalytic binding pocket.[89,90] However, experimental studies have shown that this mutation only slightly increases local cleavage efficiency and does not significantly impact viral entry.[91]

In December 2020 (12th mo.), we identified multiple LMs, two of which were later documented in two different RVs: the V1176F mutation was confirmed in the γ variant that emerged in January 2021; the Q677H mutation was confirmed in the η variant that emerged in February 2021. While the mutational effect of V1176F is not well-elucidated, the Q677H mutation of the η SD2 has been found to promote syncytia formation.[92] In addition to these novel mutations, the amino acid site 681, which was initially characterized by its high $\boldsymbol{L}$-index in April 2020, developed substantial SAP by early 2021 (Fig. 4). This was demonstrated by its substitution to various basic residues. For instance, instead of retaining α variant's P681H mutation, both the kappa (κ, B.1.617.1) and δ variants developed the novel P681R mutation. Unlike its predecessor, this mutation greatly improves furin cleavage efficiency and hence spike fusogenicity.[93,94] We identified several more novel LMs within these domains in the following months, some of which were harbored by forthcoming RVs. These include the D950N mutation of the δ S2 and the T859N mutation of the λ S2 (Fig. 4 and Fig. 5).

In October 2021 (22nd mo.), we identified additional novel post-RBD LMs. One of these sites, N679K,

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

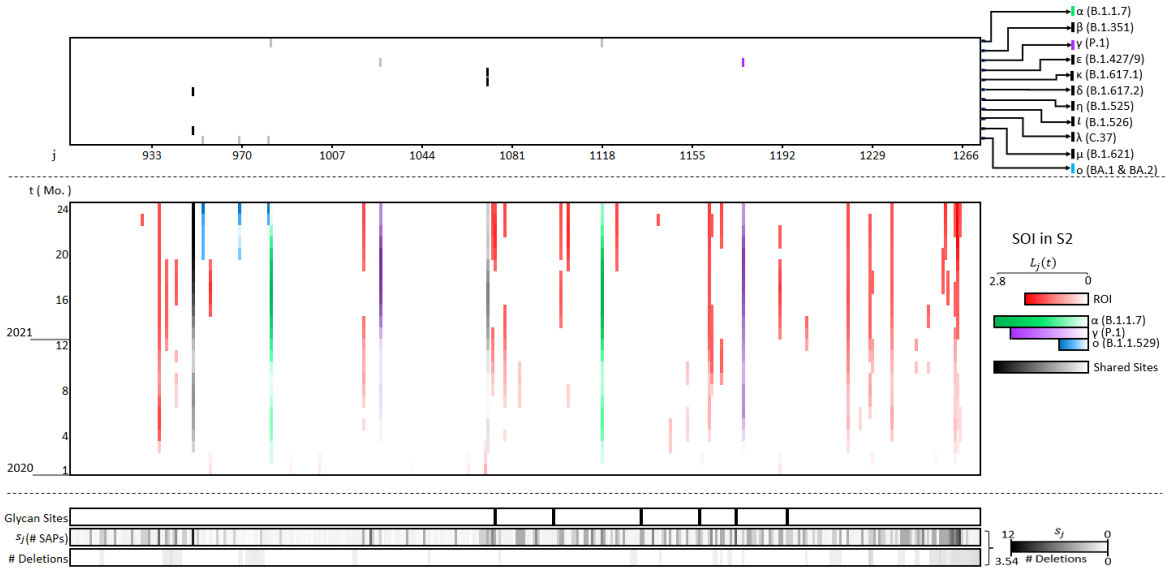

Fig. 5. Evolutionary trajectory of the S2 between January 2020 (1st mo.) and December 2021 (24th mo.). Top panel: S2 mutations harbored by RVs as defined by the WHO. Sites shared by more than 1 RV are colored in black. Middle panel: S2 LMs outlined by our deLemus analysis, whose mutational activities are measured by their $\boldsymbol{L}_j(t)$. Bottom panel: Distributions of glycosylation sites, $\boldsymbol{s}_j(t)$, and deletions in the S2, as computed by their mean values of occurrence in all non-degenerate sequences collected between March to August 2021.

was later confirmed to be present in the SD2 of the *o* lineage, shared by both the BA.1 and BA.2 strains (Fig. 4). This mutation had previously been detected in January 2021 (13th mo.). The N679K mutation introduces a basic lysine residue near the S1/S2 furin cleavage site. However, experimental studies have suggested that, like the P681H mutation of α, N679K does not appreciably improve proteolytic processing of the spike glycoprotein.[95]

Aside from the outlined LMs, we also revealed some general evolutionary features of the SARS-CoV-2 spike glycoprotein, which can be quantified by Ξ and $\bar{\boldsymbol{s}}$ respectively. Mutations are the primary source of genetic variation; how mutation rates fluctuate over the course of evolution is of particular concern.[96] In terms of its mutation rate, 3 characteristic phases of Ξ evolution can be distinguished throughout the pandemic. In the first phase between the initial outbreak of SARS-CoV-2 in December 2019 and October 2020, Ξ maintained a steady state at a relatively low value. A month before the emergences of α and β in December 2020, the second phase proceeded, when Ξ gradually increased until attaining a maximum value in March 2021, coinciding with the emergence of δ. Afterwards, the third phase followed, which marked the gradual decrease of Ξ up till June 2021, as it returned to a steady state comparable to that noted in the first phase. As most mutations on the spike glycoprotein are assumed to be deleterious, selection often acts against high mutation rates,[97] which can be illustrated by the persistently low Ξ in the first evolutionary phase. Additionally, the stable Ξ in this phase may implicate the establishment of a dynamic equilibrium within a heterogeneous viral population,[98,99] where each pre-α variant harbors similar levels of fitness. The subsequent rise in Ξ is thought to be a consequence of environmental changes that perturbed the equilibrium population of SARS-CoV-2. For better adaptability, elevating Ξ via the gain of a mutator allele would provide increased chances from which beneficial mutations would arise.[100] However, more deleterious mutations would likewise be introduced if high Ξ value remains sustained.[97,101] This would explain the gradual reduction of Ξ upon the start of the third phase. On the other hand, while genetic diversity measured by $\bar{\boldsymbol{s}}$ also exhibits slight fluctuations over the course of this pandemic, a clear increasing trend can be observed since December 2019. This growing $\bar{\boldsymbol{s}}$ indicates the ongoing adaptive diversification of the SARS-CoV-2 spike glycoprotein.[102] Nonetheless, further research would be necessary to elucidate the exact mechanisms that govern the complex evolutionary trajectory of SARS-CoV-2.

IV. Conclusion

Our deLemus analysis allows the systematic monitoring of dynamic protein sequence samples, by extracting crucial information from time-dependent big data sets that accounts for the mutational activity of each amino acid site. Based on the variant decomposition data shown in Fig. S4, we

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

defined the emergence of the α, δ, and o variants as the three most prominent phases of the COVID-19 pandemic. LMs outlined by our deLemus analysis were therefore compared against the reported mutations of these three RVs, as summarized in Fig. S8. Prior to the outbreak of the α variant, we captured seven of its ten reported mutations. Moving on to the δ variant, we were able to outline eight of its nine mutational sites as LMs. Lastly, we confirmed half of the mutational sites carried by the o variant. Overall, around 70% of the aforementioned RV mutations correspond to our outlined LMs, meaning that our deLemus analysis is capable of effectively capturing most RV mutations as LMs from the enormous data sets. Hence, these outlined LMs can serve as a guideline for deciphering the complex mutational pathways taken by the SARS-CoV-2 spike glycoprotein. Moreover, our deLemus analysis enables us to comprehend the time-resolved mutation patterns in each functional domain of the spike glycoprotein. For the NTD, our results verified the presence of previously characterized RDRs,[9] and showed that these RDRs evolve by expansion over time, as indicated by their monthly $\boldsymbol{L}$-index computations depicted in Fig. 2. This highlights the importance of deletions in the evolutionary process of SARS-CoV-2. As well, our results showed a high occurrence of SAP in the NTD, particularly in common mAb-targeted regions.[45,103] Collectively, the disproportionately high mutation rate of the NTD implies that the virus favors evolutionary routes in which both optimal ACE2-binding affinity and domain flexibility for immune escape are retained. For the RBD, we observed that most mutations harbored by previous RVs were incorporated into the RBD of the o variant, likely because these mutations are linked to fitness-enhancing traits. Given the RBD's crucial role in mediating viral infections, which already shows robust binding affinity toward ACE2 receptors, we propose that the emergence of LMs in this region is driven by immune selection pressure. As a result, RBD mutations are likely to enhance viral transmissibility and antibody resistance.[53] The rapid evolution of the SARS-CoV-2 spike glycoprotein is evident from its increasing number of mutated amino acid sites $\boldsymbol{n}$ and monthly SAPs $\bar{\boldsymbol{s}}$ over time, as suggested by our deLemus analysis. To develop better pandemic control measures, it is quintessential to first understand how SARS-CoV-2 evolves. Therefore, sustained global efforts have to be called for a more comprehensive virus surveillance system.

V. 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), ASPIRE League Partnership Seed Fund Program (ASPIRE 20191), 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 and Siqi Lei for improving figures; Ruichen Li, Leon R. Salim, and Salina Lam in data collection and preparation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. 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).
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. Abdelrahman, Z., Li, M. & Wang, X. Comparative Review of SARS-CoV-2, SARS-CoV, MERS-CoV, and Influenza A Respiratory Viruses. *Front. Immunol.* **11** (2020).
7. Liu, C., Shi, W., Becker, S. T., *et al.* Structural basis of mismatch recognition by a SARS-CoV-2 proofreading enzyme. *Science* **373,** 1142–1146 (2021).
8. Moeller, N. H., Shi, K., Özlem Demir, *et al.* Structure and dynamics of SARS-CoV-2 proofreading exoribonuclease ExoN. *Proc. Natl. Acad. Sci. U.S.A.* **119,** e2106379119 (2022).
9. 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).
10. Li, X., Giorgi, E. E., Marichannegowda, M. H., *et al.* Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv.* **6,** eabb9153 (2020).
11. Turakhia, Y., Thornlow, B., Hinrichs, A., *et al.* Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature* **609,** 994–997 (2022).
12. Simon-Loriere, E. & Holmes, E. C. Why do RNA viruses recombine? *Nat. Rev. Microbiol.* **9,** 617–626 (2011).
13. 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).
14. *Tracking SARS-CoV-2 variants.https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/*
15. Carabelli, A. M., Peacock, T. P., Thorne, L. G., *et al.* SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat. Rev. Microbiol.* (2023).
16. 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).
17. McCormick, K. D., Jacobs, J. L. & Mellors, J. W. The emerging plasticity of SARS-CoV-2. *Science* **371,** 1306–1308 (2021).
18. Jackson, C. B., Farzan, M., Chen, B., *et al.* Mechanisms of SARS-CoV-2 entry into cells. *Nat. Rev. Mol. Cell Biol.* **23,** 3–20 (2022).
19. 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).
20. 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).
21. 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).
22. 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).
23. Hansen, J., Baum, A., Pascal, K. E., *et al.* Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* **369,** 1010–1014 (2020).
24. Cohen, K. W., Linderman, S. L., Moodie, Z., *et al.* Longitudinal analysis shows durable and broad immune memory after SARS-CoV-2 infection with persisting antibody responses and memory B and T cells. *Cell Rep. Med.* **2,** 100354 (2021).
25. Moss, P. The T cell immune response against SARS-CoV-2. *Nat. Immunol.* **23,** 186–193 (2022).
26. Huang, Y., Yang, C., Xu, X.-f., *et al.* Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol. Sin.* **41,** 1141–1149 (2020).
27. Duan, L., Zheng, Q., Zhang, H., *et al.* The SARS-CoV-2 Spike Glycoprotein Biosynthesis, Structure, Function, and Antigenicity: Implications for the Design of Spike-Based Vaccine Immunogens. *Front. Immunol.* **11** (2020).
28. Dai, L. & Gao, G. F. Viral targets for vaccines against COVID-19. *Nat. Rev. Immunol.* **21,** 73–82 (2021).
29. 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).
30. 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).
31. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22,** 30494 (2017).
32. Wu, F., Zhao, S., Yu, B., *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579,** 265–269 (2020).
33. Goujon, M., McWilliam, H., Li, W., *et al.* A new bioinformatics analysis tools framework at EMBL–EBI. *Nucleic Acids Res.* **38,** W695–W699 (2010).
34. Golub, G. H. & Van Loan, C. F. *Matrix Computations* 4th ed., 780 (The Johns Hopkins University Press, Maryland, 2013).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

35. Jumper, J., Evans, R., Pritzel, A., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).
36. Wrapp, D., Wang, N., Corbett, K. S., *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367,** 1260–1263 (2020).
37. 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).
38. Wang, Z., Muecksch, F., Cho, A., *et al.* Analysis of memory B cells identifies conserved neutralizing epitopes on the N-terminal domain of variant SARS-Cov-2 spike proteins. *Immunity* **55,** 998–1012.e8 (2022).
39. Qing, E., Kicmal, T., Kumar, B., *et al.* Dynamics of SARS-CoV-2 Spike Proteins in Cell Entry: Control Elements in the Amino-Terminal Domains. *mBio* **12,** e01590–21 (2021).
40. Ray, D., Le, L. & Andricioaei, I. Distant residues modulate conformational opening in SARS-CoV-2 spike protein. *Proc. Natl. Acad. Sci. USA.* **118,** e2100943118 (2021).
41. Alisoltani, A., Jaroszewski, L., Iyer, M., *et al.* Increased Frequency of Indels in Hypervariable Regions of SARS-CoV-2 Proteins—A Possible Signature of Adaptive Selection. *Front. Genet.* **13** (2022).
42. Weng, S., Zhou, H., Ji, C., *et al.* Conserved Pattern and Potential Role of Recurrent Deletions in SARS-CoV-2 Evolution. *Microbiol. Spectr.* **10,** e02191–21 (2022).
43. 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).
44. 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).
45. 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).
46. 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).
47. 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).
48. Dhar, M. S., Marwal, R., VS, R., *et al.* Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. *Science* **374,** 995–999 (2021).
49. Chaudhari, A. M., Joshi, M., Kumar, D., *et al.* Evaluation of immune evasion in SARS-CoV-2 Delta and Omicron variants. *Computat. Struct. Biotechnol. J.* **20,** 4501–4516 (2022).
50. Kimura, I., Kosugi, Y., Wu, J., *et al.* The SARS-CoV-2 Lambda variant exhibits enhanced infectivity and immune resistance. *Cell Rep.* **38,** 110218 (2022).
51. Karim, S. S. A. & Karim, Q. A. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *Lancet* **398,** 2126–2128 (2021).
52. Viana, R., Moyo, S., Amoako, D. G., *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603,** 679–686 (2022).
53. Liu, L., Iketani, S., Guo, Y., *et al.* Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602,** 676–681 (2022).
54. Haslwanter, D., Dieterle, M. E., Wec, A. Z., *et al.* A Combination of Receptor-Binding Domain and N-Terminal Domain Neutralizing Antibodies Limits the Generation of SARS-CoV-2 Spike Neutralization-Escape Mutants. *mBio* **12,** e02473–21 (2021).
55. Tai, W., He, L., Zhang, X., *et al.* Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cell. Mol. Immunol.* **17,** 613–620 (2020).
56. 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).
57. Premkumar, L., Segovia-Chumbez, B., Jadi, R., *et al.* The receptor-binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in SARS-CoV-2 patients. *Sci. Immunol.* **5,** eabc8413 (2020).
58. Piccoli, L., Park, Y.-J., Tortorici, M. A., *et al.* Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell* **183,** 1024–1042.e21 (2020).
59. Tao, K., Tzou, P. L., Nouhin, J., *et al.* The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* **0123456789** (2021).
60. Wang, Z., Schmidt, F., Weisblum, Y., *et al.* mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature* **592,** 616–622 (2021).
61. Weisblum, Y., Schmidt, F., Zhang, F., *et al.* Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife* **9** (2020).
62. 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).
63. 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).
64. 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).
65. Saville, J. W., Mannar, D., Zhu, X., *et al.* Structural and biochemical rationale for enhanced spike protein fitness in delta and kappa SARS-CoV-2 variants. *Nat. Commun.* **13,** 742 (2022).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

66. 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).
67. Planas, D., Veyer, D., Baidaliuk, A., *et al.* Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596,** 276–280 (7871 2021).
68. 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).
69. 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).
70. Chen, J., Wang, R., Wang, M., *et al.* Mutations Strengthened SARS-CoV-2 Infectivity. *J. Mol. Biol.* **432,** 5212–5226 (2020).
71. 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).
72. Wang, Y., Liu, C., Zhang, C., *et al.* Structural basis for SARS-CoV-2 Delta variant recognition of ACE2 receptor and broadly neutralizing antibodies. *Nat. Commun.* **13,** 871 (2022).
73. Aggarwal, A., Naskar, S., Maroli, N., *et al.* Mechanistic insights into the effects of key mutations on SARS-CoV-2 RBD–ACE2 binding. *Phys. Chem. Chem. Phys.* **23,** 26451–26458 (46 2021).
74. Cao, Y., Yisimayi, A., Jian, F., *et al.* BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature* **608,** 593–602 (7923 2022).
75. 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).
76. Gobeil, S. M.-C., Janowska, K., McDowell, S., *et al.* Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. *Science* **373,** eabi6226 (2021).
77. Fan, X., Cao, D., Kong, L., *et al.* Cryo-EM analysis of the post-fusion structure of the SARS-CoV spike glycoprotein. *Nat. Commun.* **11,** 3618 (2020).
78. Henderson, R., Edwards, R. J., Mansouri, K., *et al.* Controlling the SARS-CoV-2 spike glycoprotein conformation. *Nat. Struct. Mol. Biol.* **27,** 925–933 (2020).
79. Gobeil, S. M.-C., Janowska, K., McDowell, S., *et al.* D614G Mutation Alters SARS-CoV-2 Spike Conformation and Enhances Protease Cleavage at the S1/S2 Junction. *Cell Rep.* **34,** 108630 (2021).
80. Yang, T.-J., Yu, P.-Y., Chang, Y.-C., *et al.* Effect of SARS-CoV-2 B.1.1.7 mutations on spike protein structure and function. *Nat. Struct. Mol. Biol.* **28,** 731–739 (2021).
81. Yu, S., Zheng, X., Zhou, B., *et al.* SARS-CoV-2 spike engagement of ACE2 primes S2 site cleavage and fusion initiation. *Proc. Natl. Acad. Sci. U S A* **119,** e2111199119 (2022).
82. Belouzard, S., Chu, V. C. & Whittaker, G. R. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proc. Natl. Acad. Sci. U S A* **106,** 5871–5876 (2009).
83. Rajah, M. M., Bernier, A., Buchrieser, J., *et al.* The Mechanism and Consequences of SARS-CoV-2 Spike-Mediated Fusion and Syncytia Formation. *J. Mol. Biol.* **434,** 167280 (2022).
84. Jaimes, J. A., Millet, J. K. & Whittaker, G. R. Proteolytic Cleavage of the SARS-CoV-2 Spike Protein and the Role of the Novel S1/S2 Site. *iScience* **23,** 101212 (2020).
85. Korber, B., Fischer, W. M., Gnanakaran, S., *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182,** 812–827.e19 (2020).
86. Gobeil, S. M.-C., Janowska, K., McDowell, S., *et al.* D614G Mutation Alters SARS-CoV-2 Spike Conformation and Enhances Protease Cleavage at the S1/S2 Junction. *Cell Rep.* **34** (2021).
87. Ozono, S., Zhang, Y., Ode, H., *et al.* SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nat. Commun.* **12,** 848 (2021).
88. Zhou, B., Thao, T. T. N., Hoffmann, D., *et al.* SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature* **592,** 122–127 (2021).
89. Henrich, S., Cameron, A., Bourenkov, G. P., *et al.* The crystal structure of the proprotein processing proteinase furin explains its stringent specificity. *Nat. Struct. Mol. Biol.* **10,** 520–526 (2003).
90. Vankadari, N. Structure of Furin Protease Binding to SARS-CoV-2 Spike Glycoprotein and Implications for Potential Targets and Virulence. *J. Phys. Chem. Lett.* **11,** 6655–6663 (2020).
91. Lubinski, B., Fernandes, M. H., Frazier, L., *et al.* Functional evaluation of the P681H mutation on the proteolytic activation of the SARS-CoV-2 variant B.1.1.7 (Alpha) spike. *iScience* **25,** 103589 (2022).
92. Zeng, C., Evans, J. P., Faraone, J. N., *et al.* Neutralization of SARS-CoV-2 Variants of Concern Harboring Q677H. *mBio* **12,** e02510–21 (2021).
93. Saito, A., Irie, T., Suzuki, R., *et al.* Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. *Nature* **602,** 300–306 (2022).
94. Liu, Y., Liu, J., Johnson, B. A., *et al.* Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant. *Cell Rep.* **39** (2022).
95. Zhang, J., Cai, Y., Lavine, C. L., *et al.* Structural and functional impact by SARS-CoV-2 Omicron spike mutations. *Cell Rep.* **39,** 110729 (2022).
96. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26,** 345–352 (2010).
97. Peck, K. M. & Lauring, A. S. Complexities of Viral Mutation Rates. *J. Virol.* **92,** e01031–17 (2018).
98. Domingo, E., Sabo, D., Taniguchi, T., *et al.* Nucleotide sequence heterogeneity of an RNA phage population. *Cell* **13,** 735–744 (1978).
99. Steinhauer, D. A. & Holland, J. J. Rapid Evolution of RNA Viruses. *Annu. Rev. Microbiol.* **41,** 409–431 (1987).
100. Sniegowski, P. D., Gerrish, P. J. & Lenski, R. E. Evolution of high mutation rates in experimental populations of E. coli. *Nature* **387,** 703–705 (1997).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.29.474427; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

101. Couce, A., Caudwell, L. V., Feinauer, C., *et al.* Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **114,** E9026–E9035 (2017).
102. Ellegren, H. & Galtier, N. Determinants of genetic diversity. *Nat. Rev. Genet.* **17,** 422–433 (2016).
103. 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).