A rigorous framework for detecting SARS-CoV-2 spike protein mutational ensemble from genomic and structural features

Saman Fatihi[1,2,†], Surabhi Rathore[1,2,†], Ankit Pathak[1], Deepanshi Gahlot[1,2], Mitali Mukerji[1], Nidhi Jatana[1], Lipi Thukral[1,*]
[1]CSIR-Institute of Genomics and Integrative Biology, Mathura Road, New Delhi- 110 025, India.
[2]Academy of Scientific and Innovative Research (AcSIR), CSIR-Human Resource Development Centre, (CSIR-HRDC), Kamla Nehru Nagar, Ghaziabad, Uttar Pradesh- 201002, India
[†]co-first authors
[*]Correspondence to Lipi Thukral (lipi.thukral@igib.res.in)

Abstract

The recent release of SARS-CoV-2 genomic data from several countries has provided clues into the potential antigenic drift of the coronavirus population. In particular, the genomic instability observed in the spike protein necessitates immediate action and further exploration in the context of viral-host interactions. Here we dynamically track 3,11,795 genome sequences of spike protein, which comprises 2,584 protein mutations. We reveal mutational genomic ensemble at different timing and geographies, that evolves on four distinct residues. In addition to the well-established N501 mutational cluster, we detect the presence of three novel clusters, namely A222, N439, and S477. The robust examination of structural features from 44 known cryo-EM structures showed that the virus is deploying many mutations within these clusters on structurally heterogeneous regions. One such dominant variant D614G was also simulated using molecular dynamics simulations and, as compared to wild-type, we found higher stability with human ACE2 receptor. There is also a significant overlap of mutational clusters on known epitopes, indicating putative interference with antibody binding. Thus, we propose that the resulting coaxility of mutational clusters is the most efficient feature of SARS-CoV-2 evolution and provides precise mutant combinations that can enable future vaccine re-positioning.

Keywords: Spike protein, SARS-CoV-2, mutational genomic ensemble, coaxiality, structural genomics, conformational dynamics

Introduction

The current novel coronavirus disease 2019 (COVID-19) outbreak is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In Feb 2021, almost a year after it was

declared as a global pandemic, the World Health Organisation (WHO) reported 106,991,090 confirmed cases with 2,347,015 deaths all over the world. These numbers have placed an enormous burden on the health care and economic sector. Along with rigorous testing, numerous countries accelerated their genomics surveillance programs to track the genetic diversity of the virus. The first genome sequence of SARS-CoV-2 was submitted on Jan 11, 2019, referred to as Wuhan reference sequence (Wu et al., 2020). However, the persistent presence in host populations for more than a year has led to the increased genetic diversity of SARS-CoV-2 (van Dorp et al., 2020; Choi et al., 2020).

Especially, multiple mutations are accruing on the viral receptor protein of SARS-CoV-2, spike (S) protein that interacts directly with the human ACE2 receptor to infect the host cell (Yan et al., 2020). Structurally, much of this coordination involves the receptor-binding domain (RBD) located at the top of the S protein that is present in either down state or closed conformation (Wang et al., 2020). Multiple conformational rearrangements of open positions of RBD are plausible, with three-up state linked with higher binding with ACE2 protein(Henderson et al., 2020). This is followed by events that lead to virus and host cell membrane fusion (Shang et al., 2020). Experimental studies have shown that antibodies from SARS-CoV-2 recovered patients have a high neutralizing effect with strong binding to spike protein (Long et al., 2020; Jiang et al., 2020). Around 72 vaccine candidates are currently under clinical testing or at different trial phases, and the unprecedented pace of several discoveries has led to clearance of several vaccine products from regulatory agencies (Parker et al., 2020). Till date, 160 million vaccine doses have been administered around the globe. Several attempts have been made to neutralize the SARS-CoV-2 virus in hosts and design specific antibodies to stop the spread but due to growing mutations within spike protein, the interactions with antibodies have been altered (Li et al., 2020).

In April 2020, a single amino acid substitution D614G, became a widespread dominant variant (Korber et al., 2020). It has been observed that the patients carrying the D614G mutation displayed higher upper respiratory tract viral load than the wild type but the disease severity was not altered (Korber et al., 2020). Also, enhanced viral replication was observed in human lung epithelial cells and primary human airway tissues via increasing the infectivity and stability of the virus (Plante et al., 2020). Cryo-EM structures of the D614G spike structure have revealed that the mutant D614 is stable with RBD in an "up" conformation(Yurkovetskiy et al., 2020). Also, the furin site was more susceptible to cleavage in the G614 spike (Gobeil et al., 2020). Recently, multiple co-occurring mutations in the

SARS-CoV-2, including spike protein were reported (Rees-Spear et al., 2021). Three major clades, including B.1.1.7, B.1.351, and P.1 were seen in patients from UK, South Africa, and Brazil, respectively (Tegally et al., 2020). Compared to previous variants, B.1.1.7 variant raised concern due to the presence of N501Y mutation in the RBD of the spike protein. Particularly in spike, this variant emerged with several other mutations like 69/70 del and P681H, present in the N-terminal domain (NTD) and near S1/S2 site, respectively (Rambaut et al., 2020). Another variant from South Africa (B.1.351) emerged independently with multiple mutations in the RBD (K417N and E484K) along with N501Y, however, lacked the 69/70 deletion mutation in the NTD segment (Rambaut et al., 2020). Another variant P.1, a branch of the B.1.1.28 lineage contains three mutations in the RBD: K417T, E484K, and N501Y (Faria et al., 2021). Experimental studies have ramped up to explain the properties of selected mutants but the extensive clinical characterization is time-consuming.

Based on the existing pattern of mutations, it is highly challenging to predict the trajectory of the viral evolution. Here, we have combined large-scale genomic data of SARS-CoV-2 spike protein and conformational analysis of both experimentally derived cryo-EM structures and molecular dynamics (MD) simulations. Analysis of 3,11,795 genomes revealed increased mutations on unique residues, suggesting coaxial stacking of mutations. These previously unobserved mutational clusters (A222, N439, and S477) are also linked with dynamical changes within the RBD-NTD region of the spike protein. We also observed key position-dependence of residue-based clusters with known epitopes that may directly impact antibody binding. Our study illustrates generic factors that can be studied using biological data from multiple structural and genomic resources.

Results

Overview of genomics and structural data of spike protein

In this work, we have focused on two areas. First, understanding the frequency and distribution of spike mutational events in a temporal progression. Second, determining how accumulated genomic variations in spike structures relate to its function in a putative manner.

In total, 3,11,795 genomes of SARS-CoV-2 were retrieved from the GISAID database and we also decomposed genomic diversity in terms of time and population. Figure 1 illustrates 2,584 protein mutations found in these genomes submitted over a span of one year. Of these

mutations, 786 were present in the NTD domain (14-305 residues) which covers 23% of the spike length, and 402 mutations in the RBD domain (319-541 residues) (Figure S1). We detected an over-abundance of D614G mutant present in 94.3% of genomes, followed by the high prevalence of A222V, L18F, S477N, and 6 other mutations in 5-25% of genomes (Figure 2A). A sizeable fraction of these variants were found in low-frequency (<1% genomes) and were distributed throughout the length of the protein.

Nearly all prevalent mutations occur on the S1 subunit which is extensively characterised experimentally. In order to assess if these mutations occur on segments that are structurally heterogeneous, a total number of 44 cryo-EM structures corresponding to different conformational states of the spike were analysed (see Methods for PDB ids). The representative structures of each conformation are shown in (Figure 2B). Amongst these states, closed spike conformers showed large structural changes, with an average deviation of 2.77 ± 0.42 Å (Figure S2). The excluded regions with missing co-ordinates were not taken into account. High fluctuations originate within the NTD (110-187, 212-216, 242-264 a.a) and RBD (365-372, 444-447, 454-462, 468-489 a.a.). In addition, loops following RBD (542-684 a.a.) and the linker beta-sheet (306-318 a.a.) connecting the NTD and the RBD region also depicted reasonable heterogeneity. The structural overlays of these different conformations, including the variable segments in the bound forms with differential RBD opening are shown in Figure S3.

We can derive two main insights from the experimentally derived conformational ensemble of spike structures. The first is that the appearance of mutated side chains on structurally changing regions varies with different functional states. Almost 50% highly prevalent mutations occur in heterogeneous regions in closed conformers, as opposed to 33% in open positions. The distinguishability is also contributed from two crucial RBM residues (N439 and N501). The second is that large genomic variations are present within the NTD region. In a pool of 266 mutations, 163 were mapped on heterogeneous segments within the closed conformers. Our findings, thus, reveal the mutations that are involved in settings that initiate spike protein activation and overcome the limited view obtained from mapping on a single structure.

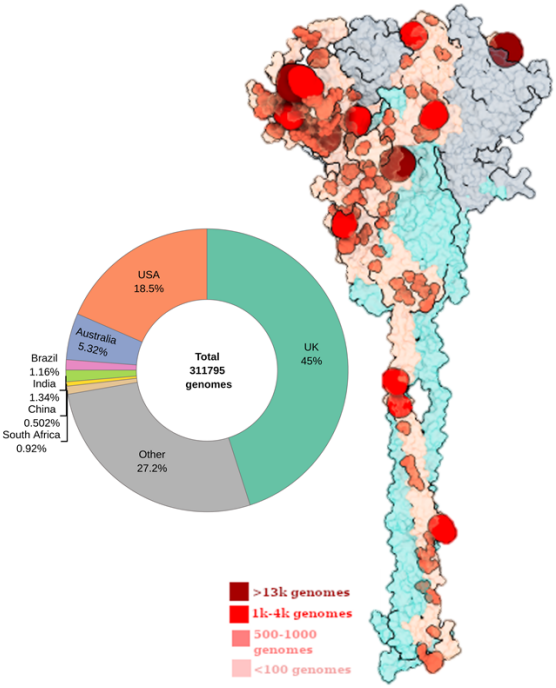

Figure 1: **Mutational spectrum of SARS-CoV-2 spike protein.** In total, 3,11,795 genomes were retrieved from GISAID, as ploted in the pie chart, with different colors showing country-wise submitted genomes. The full-length spike protein is shown in surface representation, with the S1 and S2 subunit colored in grey and cyan, respectively. In total, 2584 mutations are mapped on one chain, where the S1 and the S2 subunit depicts 1509 and 1075 amino-acid substitutions, respectively. The size of the sphere indicates frequency of mutations, with dark to low gradient showing a higher to lower frequency range.

Tracking of SARS-CoV-2 genomes across time shows enrichment of co-occurring spike mutations

As mutations occur over time, we investigated the temporal progression of spike genomic variations. Figure 2C shows the distribution of mutations per sample for the entire one-year duration, depicting a sharp increase from roughly one mutation in the early pandemic phase to ≈6 co-occurring mutations in Dec 2020. We defined mutational genomic ensemble as a pool of co-occurring spike mutants with different residue combinations present in at least 0.5% of the total samples submitted across the 14 day period. As shown in the heatmap, this revealed an array of co-occurring spike mutants at a different time intervals. At the start of the pandemic, the initial Wuhan sequence represented a major part of the total genomes and did not display any mutation comprising 6.3% of the screened genomes. However, only D614G dominated and persisted throughout the time, covering 94.3% of the submitted genomes. In particular, other residues appeared in combination with D614G that persisted over time, namely P1263L, D839Y, D936Y, Q675H, V1176F, and H49Y. We

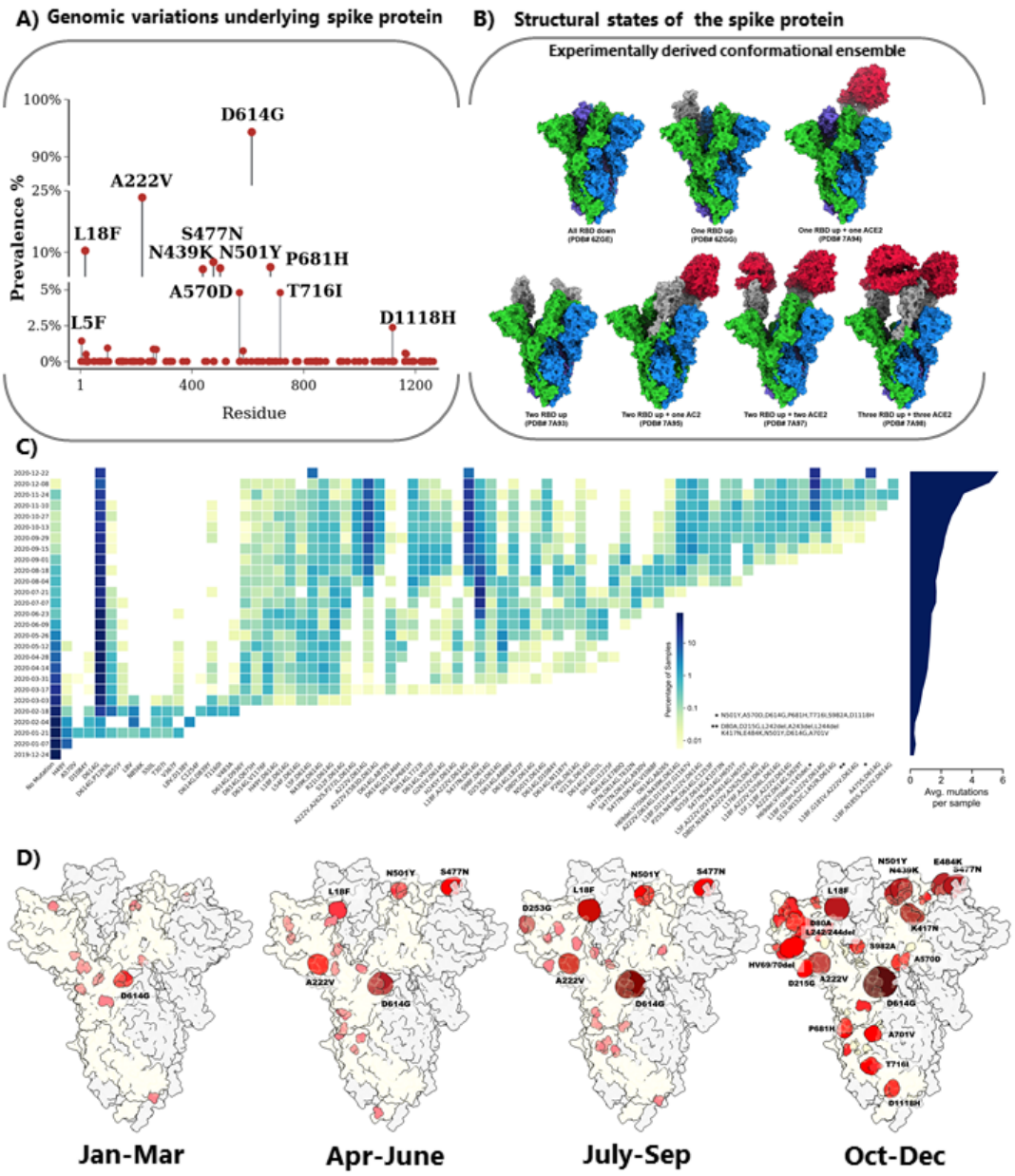

Figure 2: **Dynamical tracking of genomes reveal spike protein mutational ensemble.** A) The needle plot shows the residue-wise global frequency of mutations in the spike protein, with Y-axis showing the prevalence. Other than D614G occurring at 94.3%, two major frequency groups of 5-10% and 25-90% prevalence can be observed. B) Structural snapshots of different conformational states of the spike protein where chains A, B, and C are colored in blue, purple, and green, respectively. The up-state of RBD is shown in grey and the bound ACE2 protein is colored in red. C) The heat map shows the temporal progression of spike mutations prevalent in the 14-day period. Each row shows the allelic frequency in 0.5% of the total samples submitted. The area graph (extreme right) plots the average number of spike mutations per sample in the SARS-CoV-2 genomes as a function of time. The cumulative genomes submitted from Australia, Brazil, China, India, South Africa, United Kingdom, and the USA were analysed. D) Time progression of the co-occurring mutations illustrated on the spike structure where mutation sites are depicted in sphere representation and the sphere size indicates the prevalence of the mutation. The involvement of a residue in the formation of the number of unique combinations is denoted by the gradient of red where dark red indicates the highest number of combinations.

were particularly interested in the residues that participate in multiple residue combinations. Figure 2D shows the structural glimpse of the frequency of mutation and presence of that particular residue as a co-occurring mutation across four prominent three-month time blocks. The initial snapshot indicates the presence of several mutations that appeared in the Jan-March 2020 period along with D614G, including H655Y, L8V, N856K, and S50L. Interestingly, most single mutations involved amino acid substitutions into a hydrophobic residue, including L5F, V367F, V483A, and H49Y. Some combinations with D614G appeared in March and have retained for several months, representing 25.16% of the total genomes. However, many of the double mutants dropped as a function of time, including L8V, D1138Y which appeared for one month. The third time-quarter depicted a clear trend with two combinations displaying relatively higher prevalence. The A222V-D614G and S477N-D614G combinations, represented 6.33% and 5.2% of the total genomes, respectively. Similarly, for triple co-occurring mutations, L18F-A222V-D614G (11.9%) showed the highest prevalence. Besides this combination, there is a significant pool of low-frequency triple co-occurring mutations, namely- A222V-E583D-D614G (0.47%), S477N-D614G-T632N (0.29%), S477N-D614G-A930V (0.27%), S477N-D614G-V1068F (0.29%), S255F-D614G-K1073N, S477N-D614G-H655Y, L176F-A222V-D614G, and A222V-D614G-S929T. In July, for the first time two deletion sites H69del and V70del occurred with the N439K-D614G combination.

Later, a sudden surge in multiple co-occurring mutations was observed, with 1.08% of the genomes sequenced during this time showed more than 4 spike mutations per sample. In particular, two prevalent combinations were striking, representing now known B.1.1.7 and B.1.1.5 lineage, with 9.7% of genomes showing the former that later increased to 27% of total genomes (Rambaut et al., 2020). Interestingly in the October-December duration, 51% genomes consisted of multiple combinations of mutations involving >5 residues. From the atomistic perspective, structural representatives of this time block showed mutations accumulating on several residues. These co-occurring mutations are located in two domains in the S1 subunit, with 32 mutations distributed in NTD and 9 in the RBD region with high-frequency. Our analysis has significant advantages in understanding mutations as a function of allelic frequency per residue. Most notably, the concurrent structural maps of these combinatorial interactions, enable a more accurate measure of mutation rate, while providing clues into time-dependent interactions of co-occurring mutations.

Four distinct clusters reveal coaxility within spike protein evolution

Incrementally accruing mutations on the spike protein were observed on four residues, namely A222, N501, S477, and N439 (Figure 3). This observation was captured by tracking the occurrence of residue combinations that persisted over a given time per sample. The first resulting mutational ensembles from genomic data is A222 cluster, with 16 persistent co-occurring combinations containing 1685 unique residue combinations. The evolution of multiple spike mutations was also investigated population-wise. Several countries showed simultaneous evolution of L18F and A222V that later merged with other residues. The most prevalent combination is A222V, L18F, and D614G. The last observed combination in this cluster accounted for 5-residue combination (D80Y, N164T, A222V, A262S, and D614G).

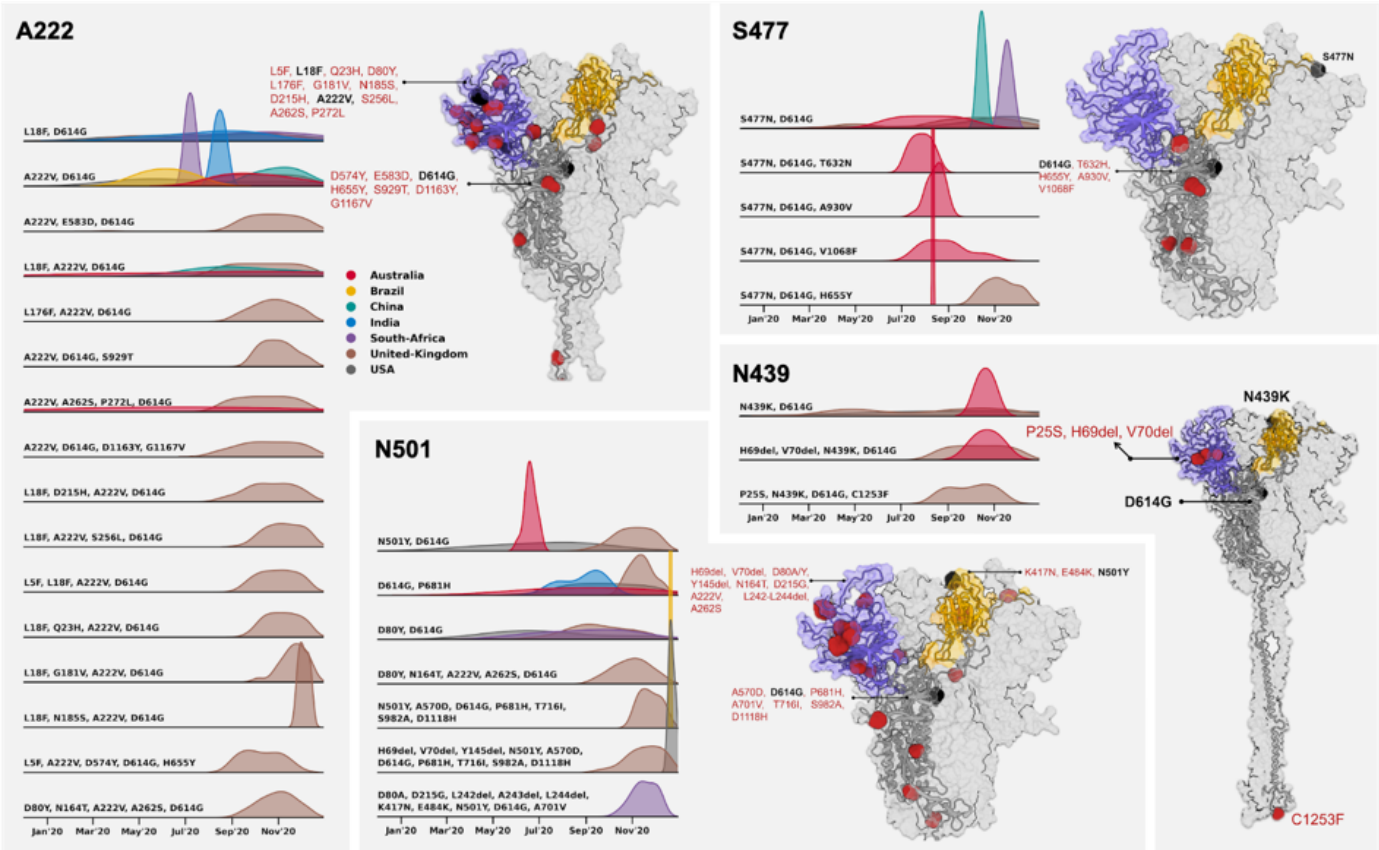

Figure 3: **Spike mutations are accruing residue-wise representing coaxiality.** Four residue-wise clusters were obtained from the mutational genomic ensemble. Ridge-plot of the density of spike mutations in genomes submitted from various populations shows the prevalence of different mutation combinations over the timeline of the pandemic. The range of the y-axis is 0 to 0.1 for each subplot (hidden for clarity). The colors of the ridges in red, yellow, green, blue, purple, brown, and grey indicate populations of Australia, Brazil, China, India, South Africa, UK, and USA respectively. The mutations are grouped as derivatives of different residues (A222, N501, S477, and N439) and are arranged in increasing order of mutation combinations. The structural mutant models were generated and the mapping of all residues involved in each cluster are marked in four representative snapshots.

The second cluster of N501 residue was abruptly observed in UK and now defined as B.1.1.7 lineage, with multiple co-occurring mutations. Several theories are proposed for these large number of genetic changes, including two deletion sites (Rambaut et al., 2020). The combination is defined by multiple spike mutations namely, HV69/70 del, Y145del, N501Y, A570D, D614G, P618H, T716I, S982A, and D1118H originated in early September and was observed first in the United Kingdom and later in the USA population. The cluster of mutations around N501 also appeared in South Africa (B.1.351) which consists of 3 deletions (L242del, A243del and L244del) and D80A, D215G, K417N, E484K, N501Y, D614G, and A701V mutations. The other two clusters-N439 and S477, are less prevalent but these mutations occur on critical regions important for ACE2 binding. The S477 cluster was observed in Australia and was found to be evolving rapidly in the UK population. The above analyses revealed critical substitutions that showed less prevalence before September 2020 but became more common as the pandemic spread. The frequency of the multiple mutations on these residues is dependent on the coverage of country-wise genomics surveillance programs. The above analysis underscores the importance of intermediate variants and point implications for other residues that may evolve as independent new combinations.

Are there regions on the spike that are varying arranged in a linear or 3D assembly that is under selective pressure to mutate? To answer, we generated eight spike mutant models with prevalent co-occurring mutations representing the two higher order combinations from each cluster (Figure S4). For the N501 cluster, deletion mutations were generated using Coot (Emsley et al., 2010), while other mutations were generated using FoldX (Schymkowitz et al., 2005). In cluster A222, 65% of the mutation sites are nestled in the NTD region, whereas for cluster N501, 54.5% of the total mutation sites are located here. Further, we observed that apart from the cluster A222, all other clusters have at least one critical mutation site in the RBD region. Comparison of relative surface area (RSA) with wild-type (WT) amino acids revealed an increase in the accessibility of 62.5% of the residues in co-occurring mutational combinations except for a few (Figure S5). D614G displayed the highest increase in RSA in all the mutant combinations, as also shown previously (Gobeil et al., 2020). Amongst all four clusters, four residues are present in the RBM motif (417, 477, 484, and 501; Figure S6). In terms of accessibility, 477 and 501 were found to display high accessibility in both wild and mutant structures except for residue K417N. On the other hand, E484 upon mutation displayed an increase in accessibility. Our structure-based mutant models showed an increase in surface accessibility gained by mutant structures, indicating it as an important regulator

linked to prevalent mutants. Taken together, these findings propose residue-wise coaxial accumulation of co-occurring mutations with mutational hotspots gathered around the NTD and RBD domains.

Dynamical changes in mutant spike protein show higher stability

A clear emergence of D614G as a major hub, with further mutations emerging from the hub D614G node, with distinguishing prevalence was observed (Figure 4A). The network diagram of mutations of the spike protein, with the node representing a residue and the edges depicts the number of connections from each participating node in multiple co-occurring combinations is shown. The size of each node is directly proportional to the involvement in a number of combinations, and the color of the node represents RBD and non-RBD residues. Interestingly, the prevalent RBD mutation sites, namely, N501Y, N439K, S477N, E484K, and K417N were observed with increase in number of unique combinations. Next, A222V and L18F appear in more unique combinations as compared to other mutation combinations.

Since D614G mutation is ubiquitous in all mutations, we constructed membrane-embedded models of wild-type and mutant structures in the atomistic details (see Methods). This allowed us to understand the relative dynamics of human receptor ACE2 with spike mutant structure (Figure 4B). To capture the conformational changes, we performed large-scale simulations of both systems for a simulated time of 450 ns, each. Our analysis revealed increased stability of the mutant protein (0.89 $\pm$ 0.007 nm) compared to the WT spike system (Figure S7). In terms of domains, high mobility was found in the CTD in both the structures, which is known to be highly mobile. Interestingly, a relative decrease in the mutant NTD (0.63 $\pm$ 0.002 nm) and RBD regions (0.74 $\pm$ 0.005 nm) was found, as shown in Figure 4C. It has been previously shown that the D614G mutation reduces the flexibility of the RBD domain (Ray et al., 2021). Followed by the RBD, the SD1 (542-591 a.a.) and SD2 (592-681 a.a.) regions showed a similar trend in both the WT and D614G systems with an average RMSD of 0.61 $\pm$ 0.003 nm. Figure 4D shows the representing snapshots highlighting dynamical changes within the NTD and RBD regions. Specifically, the RBM in D614G (0.72 $\pm$ 0.002 nm) is observed to be relatively stable as compared to that of the WT system (1.00 $\pm$ 0.009 nm).

Interestingly, each of the spike protein monomers is inter-twinned in such a fashion that

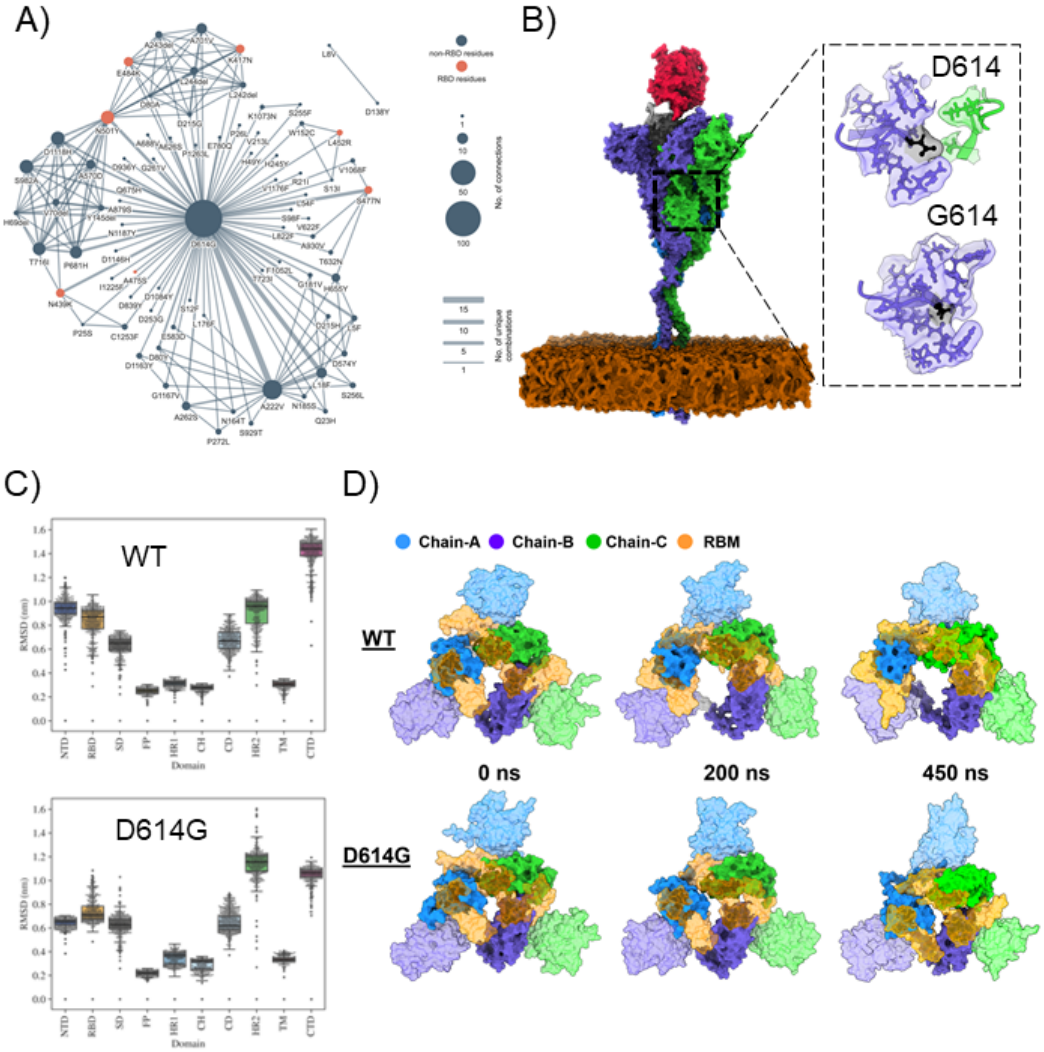

Figure 4: **Dynamics of wild-type and D614G mutant spike.** A) The graph showcases the network of co-occurring spike mutations in SARS-CoV-2 genomes where the nodes represent residues and the edges connect co-occurring mutations. The RBD residues are colored in orange while non-RBD residues in gray. The node sizes depict the number of co-occurring mutations with the residue whereas the width of the edges denote the number of unique combinations of mutations the pair of residues occur in. B) Full-length spike model in POPC membrane and bound to ACE2 is depicted. The zoomed-in image highlights the residues within 5Å of the 614th residue in WT and mutant, where 614 is highlighted in black color. C) Box plot showing the domain-wise distribution of RMSD in WT & mutant trajectories in the upper and lower panel, respectively. D) Snapshots showing the dynamics of RBM with respect to RBD and NTD in WT & mutant spike protein as a function of time. The structures are shown in surface representation with chains A, B, and C colored in blue, purple, and green, respectively while brown shows the RBM motif.

it orients the RBD closer to the next chain NTD. This results in significant cross-talk and dynamical changes interdependent from both these regions. Previous studies have shown that the mutation at the 614 site influences the RBD conformation (Yurkovetskiy et al., 2020). To check this, we analysed the correlation between the RBD and NTD of each chain

and our analyses revealed that the distances between these domains are higher in the wild-type (6.20 nm and 6.48 nm). In D614G trajectories, it corresponded to a relatively compact state with significantly less distance between the NTD and RBD (5.97 nm and 5.82 nm, respectively, as shown in Figure S8. This results in a compact conformation of the D614G and more importantly that results in higher proximity with the ACE2 structure.

RBD harbours multiple mutations linked to dynamics

Approximately 45.6% of the total mutations of the spike are present in the NTD and RBD domain, that is on the critical functional interface for human receptor and antibodies ((Chi et al., 2020; Lan et al., 2020)). Figure 5A shows the needleplot of the allelic frequency of NTD and RBD residues. The NTD has a total of 786 mutations, with A222V and L18F mutating at relatively high frequency. In total, the RBD harbors 402 mutations. Of these, 177 occur at unique amino acid sites. Except K417N, all frequently occurring mutations in the RBD, specifically occur within in RBM motif namely- S477N, N439K, N501Y, Y453F, E484K (Figure 5B).

Mapping of the flexibility parameter, root mean square fluctuations (RMSF) of both domains revealed significant stability in mutant structures (Figure S9). In wild-type, we observed distinct patches within the NTD (142-152, 180-184, and 248-257) and RBD (474 to 490, 496 to 503) that are highly mobile. This is interesting as the two prevalent mutations from the A222 cluster; A222V and L18F lie in the NTD domain.

The structural snapshots provided an atomistic view of genomic variations and dynamic regions in detail (Figure 5B). The majority of the frequently occurring RBD mutations occur on dynamic regions in the 1-up conformation of the RBD which is bound to ACE2 in the wild type. These structural maps showed that mutations such as E484K and S447N exhibit maximum flexibility in wild type with a RMSF value of 1.06 and 1.05 nm, respectively. Similarly, N501Y and N439K show a similar trend of increased flexibility in the WT. However, A520S which is present in the loop has very low flexibility in both wild and mutant. Overall the up-RBD bound to ACE2 receptor shows a striking difference in the flexibility between wild and D614G mutant structure, where the mutant is structurally more stable.

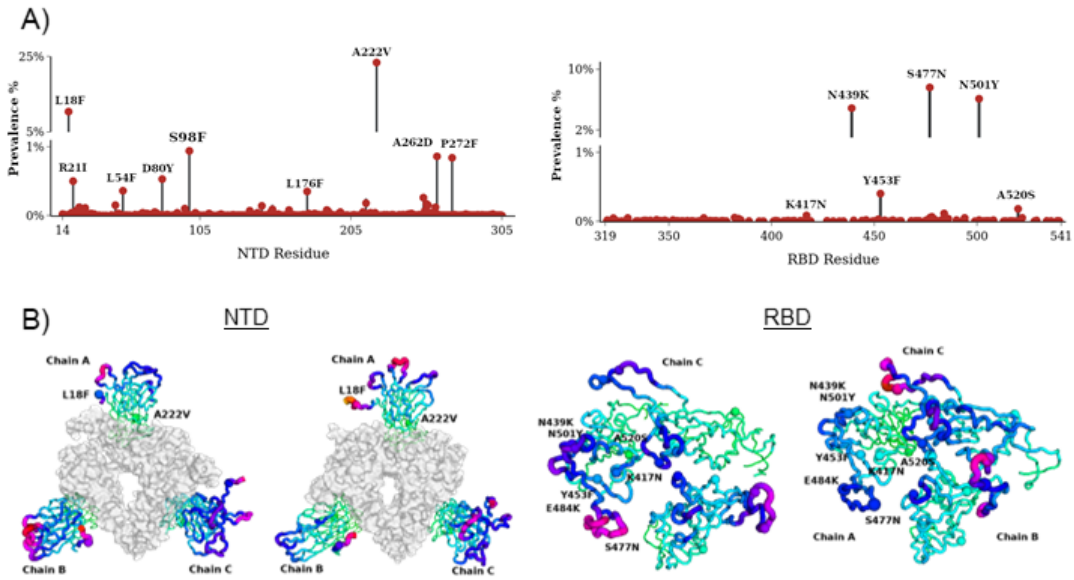

Figure 5: **Comparative dynamics & flexible region within NTD and RBD in wild-type and D614G mutant spike.** A) Needleplot showing the global frequency of mutations in the NTD (14-305) and RBD (319-541) segments. B) Snapshots displaying the RMSF in NTD and RBD domain of wild type and D614G mutant, respectively highlighting the flexible regions. The wider and red color tube indicates high fluctuating regions while shades of blue and a narrower tube signify less flexibility. Mutations present in these regions are shown in sphere representation and also marked in the text.

Mutational clusters may be linked to escape variants

Various surface epitopes have been proposed to be present on the spike structures, as shown in Figure 6A. Three distinct regions, namely, N-terminus (residues 14-20), a supersite beta-hairpin (residues 140-158), and a supersite loop (residues 245-264) constitute epitopes within the NTD region (McCallum et al., 2021). Similarly, the RBD epitopes are ridge (residues 455-456, 471-491), loop (residues 443-453, 494-502), and core (residues 365-373, 382-387) (Greaney et al., 2021). We have mapped the occurrences of mutations on these known epitopes (Figure 6B-E).

Remarkably, 22 mutations within the N501 cluster were mapping directly on the known surface epitopes. Out of 46% mutations present in NTD sites, 14% of the mutations were present on the epitopes whereas 32% of mutations were present in the proximity of the epitopes. This cluster harbors mutations on critical residues N501Y, E484K, K417N, which are known to interact with ACE2 (Xie et al., 2021). One of the most prevalent mutation "D614G" showed increased interaction with ACE2, however, there was no distinguishability in antibody binding (Weissman et al., 2021). Similarly, the N501Y mutation with high

transmissibility showed a similar pattern of strong interactions with ACE2 and no change in antibody binding (Xie et al., 2021). The spread of A222 cluster is not significant yet, however, it has multiple residues around the epitopes. In particular, this cluster containing the 20 mutations, exhibited 25% of the total mutations to be present in the NTD region. We observed 15% mutations on the epitopes and 10% mutations were present around the periphery of epitopes. Similarly, the S477 and N439 clusters, which had a total of six mutations each, displayed 16% and 66% of mutations in the NTD or RBD region, respectively. In the N439 cluster, we didn't observe any mutations on the epitope but four out of six were present around the epitopes.

The surface of the spike protein is heavily glycosylated. There are 22 N-glycosylation and O-glycosylation sites (Watanabe et al., 2020; Shajahan et al., 2020). To understand if mutations are interfering with glycosylation sites, we mapped the mutational clusters on the structure provided by Casalino et. al (Casalino et al., 2020). This detailed fully-glycosylated spike structure enabled us to study if the glycans shield any of the mutational cluster. We observed that there is no mutation that is directly present on the glycosylated residues, however, the co-occurring mutations are near to the glycosylated residues in either linear or 3D manner. In A222 cluster, the mutations L18F, N164T, D614G, H655Y, D1163Y were present near the glycosylation site N17, N165, N616, N657, N1158, respectively. The S477 cluster had mutations D614G, H655Y, V1068F near the 616, 657, and 1074 glycosylation sites, respectively. The N439 cluster had mutations at H69del, V70del, D614G which were present near the N74 and N616 glycosylation sites, respectively. In the N501 cluster, the mutations H69del, V70del, D80A/Y, N164T, D614G, T716I were present near N74, N165, D614G, N717 glycosylation sites, respectively. The impact of particular residues N501Y, E484K, L18F on the binding affinity of spike with ACE2 and antibodies have been studied but the mutations present in the proximity of critical glycosylated residues and their impact on the antibody binding remains to be studied (McCallum et al., 2021; Weisblum et al., 2020) .

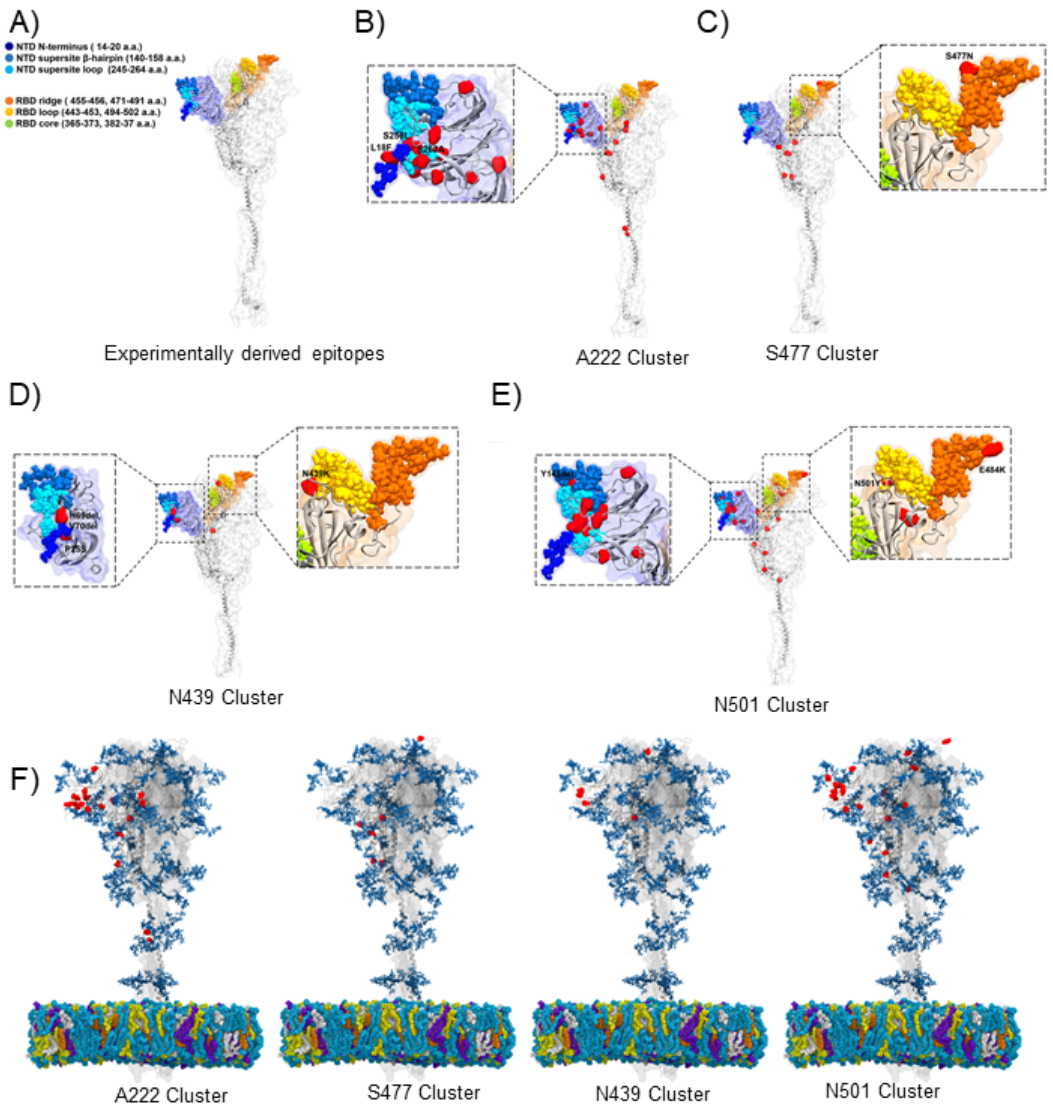

Figure 6: **Mutational clusters and epitope profiling.** A) Snapshot of full-length spike structure illustrating epitopes on NTD and RBD segments highlighted in shades of blue, and orange, respectively. B-E) Mapping of A222, S477, N439, and N501 cluster, respectively along with the NTD and RBD epitopes. The mutations present within or around the NTD and RBD epitopes are shown in the enlarged view. F) The fully glycosylated closed spike model structure was derived from Casalino et al. dataset. The complex model is embedded in the heterogeneous lipid membrane composed of POPC, POPI, POPS, POPE and cholesterol. Each spike monomer contains 22 glycosylations, shown in blue colour. Mapping of four mutational clusters derived from our work are highlighted in red spheres.

Discussion and conclusions

There are several driving forces for and against the selection determining the fitness advantage. The timely tracking of the sequencing of SARS-CoV-2 genomes in parallel with a large number of structural data has enabled several quick biological discoveries (Neher et al., 2020). The genomes served as quick indicator maps and provided real-time tracking of key mutations (Rambaut et al., 2020; Faria et al., 2021). Recently, multiple spike mutations from different geographies were reported, suggesting rapid convergent evolution (Phelan et al., 2020; Patiño-Galindo et al., 2020).

Here, we quantitatively evaluated the genomic mutational ensemble of spike protein that depicts somewhat "steady-state average" for a particular time period. Dynamical tracking of variations submitted in GISAID over a span of a year enabled a pan-spike mutational analysis that revealed more than 2500 mutations on the protein. Figure 7 shows the progression of major events involving spike protein mutations showing the accumulation of mutations. The submission of SARS-CoV-2 whole genome in Jan 2020 propelled a series of genomic tracking, with D614G identified as a major mutant associated with high transmissibility (Korber et al., 2020). In contrast to the first few months of the current pandemic that revealed 98% of the viral genome as mutation free, the mutation rate towards the end of 2020 appeared rather rapid.

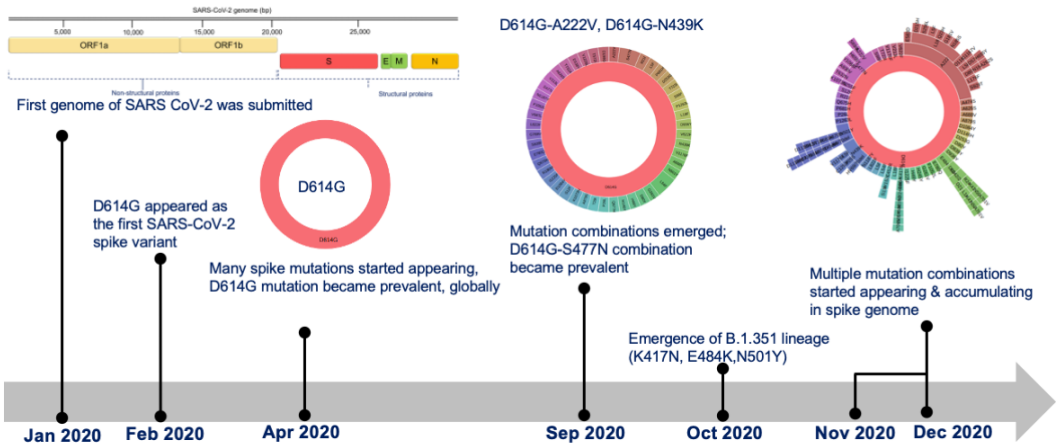

Figure 7: **Timeline of the SARS-CoV-2 spike variants emergence and accumulation.** The time progression depicts key events related to mutational events of the spike protein. The first event of genome annotation of SARS-CoV-2 was the first to properl further genomics studies. The sunburst graphs displays the occurrence of prevalent mutation D614G in March followed by single amino-acid substitutions with D614G over time. Lastly, multiple key mutations within spike RBD alongwith other combinations in various lineages were reported.

The mutational ensemble revealed spike mutation accumulation on specific residues (N501, A222, N439, and S477) as mutational hotspots that are accruing multiple mutations in dif-

ferent combinations, indicating coaxial framework of mutations. Analysis of structure-based conformational ensemble revealed that the appearance of mutated side chains on structurally heterogeneous regions varies with different functional states. Using a variety of quantitative methods, we demonstrated for the first time, the time-dependent evolution of four distinct clusters, suggesting a putative change in key functional regions within spike. Clearly, much remains to be learned, in terms of which mutational combination changes the biological function of the virus. Either way, analysis of high-throughput data and association with clinical markers must go hand in hand.

Materials and Methods

Genome data acquisition and analysis

In total, 3,11,795 spike protein sequences of SARS-CoV-2 genome were downloaded from the GISAID database submitted till 3 January, 2021 (Shu & McCauley, 2017). The low-quality sequences were removed, with the genomes having high coverage greater than 90% coverage were considered. 2,75,629 genomes passed the filtering criteria and were taken for further analysis. These SARS-CoV-2 spike genome sequences also contain geography and date of submission, which was considered for this study. Specifically, we selected 7 geographic regions, China (1,434), India (4,006), Brazil (1064), South Africa (2617), Australia (1,3112), USA (50,582), and from the United Kingdom (1,06,639); constituting 72.8% of the total genomes. Any sequences without proper submission details were dropped out. The final set of selected sequences underwent sequence analysis to determine variations in the coding region for spike protein. This was done using CoVsurver enabled by GISAID in GISAID's EpiCoV database and Wuhan 2019-nCov genome (GISAID Accession ID: EPI_ISL_402124) was used as the reference genome for mutational analysis (Okada et al., 2020). The quality check was performed and all the abrupt genomic sequences with more than 20 mutations per sample were removed. The prevalence rate for each of the observed mutations was calculated as the frequency of the mutation in total number of genomes. The mutations appearing together per spike protein sequence were annotated as co-occurring mutation combinations. To capture the time progression of these co-occurring spike mutations, we analyzed sequences submitted across the 14-day period to filter mutations present in at least 0.5% of the total samples. Additionally, these co-occurring mutation combinations were divided into four monthly quarters based on their emergence and prevalence in each of these time blocks and were mapped onto the spike structure.

Structural ensemble of experimental spike structures and analysis

A total number of 44 cryo-EM structures for various spike conformations; namely - closed unbound (all RBD down) state, partially open apo- (1 RBD-up and 2 RBD-up) states, ACE2 bound-partially opened (1 RBD-up/2 RBD-up bound to ACE2) states and active up bound to ACE2) states were taken from PDB database. 18 cryo-EM structures for

closed spike conformation (PDB ID: 6ZGE, 6VXX, 6X2C, 6X6P, 6X29, 6X79, 6XF5, 6XLU, 6XM5, 6ZB4, 6ZB5, 6ZGI, 6ZP0, 7CAB, 7DDD, 7DF3, 7JJI and 7JWY) were taken from RCSB PDB database. For 1 RBD-up (apo-state), 9 structures were taken, namely- (PDB ID: 6ZGG, 6VYB, 6X2A, 6XF6, 6XM0, 6XM3, 6XM4, 6Z97 and 7KJ5), whereas 5 cryo-EM structures were available for 1 RBD-up bound to ACE2 conformation (PDB ID: 7A94, 7DF4, 7KJ2, 7KNB and 7KNE). Similarly, for 2 RBD up (apo-state); 2 PDBs (PDB ID: 7A93 and 6X2B), for 2 RBD-up bound to 1 ACE2; 2 PDBs (PDB ID: 7A95 and 7A96) and 4 PDBS for 2 RBD-up bound to 2 ACE2 conformation (PDB ID: 7A97, 7KJ3, 7KMZ and 7KNH) were taken. For the completely open 3 RBD-up bound to ACE2 state, 4 cryo-EM structures are available (PDB ID: 7A98, 7KJ4, 7KMS, and 7KNI). The closed unbound state was taken as the reference to analyze and capture the specific domain-orientations within the representatives of other conformers of the spike protein.

To capture the structural variability within the conformers from each class of spike state, we carried out a structural alignment analysis. The analysis was done using the MultiSeq plugin for VMD software (Roberts et al., 2006). Multiseq employs the STAMP (Structural Alignment of Multiple Proteins) tool for the structural alignment and its algorithm minimizes the C-alpha distance between aligned residues of each molecule by applying globally optimal rigid-body rotations and translations. The alignment score per residue is assigned in terms of Qres, where Q refers to the fraction of similar native contacts between the aligned residues in two proteins. When Q = 1, it indicates that the structures are identical, whereas when Q has a low score, it means the structures do not align well.

Full-length wild-type and mutant spike structural models

The structural model of full-length closed spike protein was generated using the cryo-EM structure for all RBD down conformation (PDB ID: 6VXX) as template in Modeller v9.21 (Webb & Sali, 2016). In addition, the HR2 region was modeled using the NMR structure of the HR2 domain from SARS-CoV (PDB ID: 2FXP) while that of the TM region was generated using the NMR structure of HIV-1 Env polyprotein (PDB ID: 5JYN) having a homology of 100% and 29% with the spike protein, respectively. The missing residues in the structure were modeled as loops except for residues 1148-1161 which were restraint to form an α-helix. The structural model was generated in 3 segments: 14-1147, 1148-1213, and 1214-1273 and were joined together using the standard peptide-bond criteria in Chimera (Pettersen et al., 2004a).

In addition, eight structural mutant models with co-occurring mutations in four mutational clusters were generated. Each chain was mutated for point substitution mutations using FoldX (Schymkowitz et al., 2005), whereas deletion mutation in each chain was introduced employing Coot (Emsley et al., 2010).The solvent-accessible surface area of each residue was calculated by Naccess (Hubbard & Thornton, 1993).

Molecular Dynamics Simulations

Generation of ACE-bound 1 RBD-up spike conformation. The full-length model of spike in 1 RBD-up state was generated as above using cryo-EM structure of the spike 1 RBD-up state (PDB ID: 6VSB) as template. To obtain a complete atomistic model of ACE2 bound structures, we took the X-ray structure of human ACE2 bound to the RBD domain (PDB ID: 6M0J) to model the bound RBM segment and ACE2 ecto-domain.

System set-up and details of simulation protocol. The full-length spike model of 1 RBD-up state was oriented in the membrane containing 2068 POPC lipids. The membrane was generated using VMD and the spike structure was embedded in the membrane using Inflate-GRO2 (Humphrey et al., 1996; Kandt et al., 2007). The transmembrane residues (1214 to 1273) were positioned in the membrane by translating the protein along the Z-axis. The protein was placed in a rectangular box large enough to contain the protein and the membrane. Na+ ions were added to neutralize and TIP3P water representation was used to solvate the system (Mark & Nilsson, 2001). The simulations were carried out using GROMACS version 2018.3 by employing CHARMM36 all-atom force field (Abraham et al., 2015). Periodic boundary conditions were used and 1.2 nm was set as real space cut-off distance. Particle Mesh Ewald (PME) summation using the grid spacing of 0.16 nm was used combining with a fourth-order cubic interpolation to deduce the forces and potential in-between grid points (Darden et al., 1993). The van der Waals cut-off was set to 1.2 nm. A 2 fs time step for numerical integration of the equations of motion was used and the coordinates were saved at every 100 ps. The initial systems were subjected to energy minimization using the steepest descent method. Temperature and pressure were maintained at 310 K and 1 bar using Nose-Hoover thermostat and Parrinello-Rahman barostat, respectively (Bussi et al., 2007; Parrinello & Rahman, 1981). Simulations of both wild-type and D614G systems were carried out for 450 ns each.

All the molecular images were rendered using UCSF Chimera and ChimeraX (Pettersen et al., 2004b; Pettersen et al., 2021), VMD (Humphrey et al., 1996) and PyMOL (Schrödinger,

LLC, 2015). The graphs and plots were generated using Python libraries.

Data availability

The data files related to the genomic analyses, structural models and MD simulation trajectories have been made available at
https://github.com/CSB-Thukral-Lab/spike_genomic_and_structural_work.

Acknowledgements

This work was supported by funding from CSIR and the Department of Science and Technology (DST). S.F. is supported by DST Early Career Award, S.R. by DST-Inspire fellowship, D.G. by CSIR-NET fellowship, and N.J. by DST-SERB fellowship. We acknowledge the support from CSIR-IGIB for infrastructure and CSIR-4PI for supercomputing facilities.

References

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., & Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in china. *Nature*, *579*(7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3

van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., Owen, C. J., Pang, J., Tan, C. C. S., Boshier, F. A. T., Ortiz, A. T., & Balloux, F. (2020). Emergence of genomic diversity and recurrent mutations in sars-cov-2, 104351. https://doi.org/10.1016/j.meegid.2020.104351

Choi, B., Choudhary, M. C., Regan, J., Sparks, J. A., Padera, R. F., Qiu, X., Solomon, I. H., Kuo, H.-H., Boucau, J., Bowman, K., Adhikari, U. D., Winkler, M. L., Mueller, A. A., Hsu, T. Y.-T., Desjardins, M., Baden, L. R., Chan, B. T., Walker, B. D., Lichterfeld, M., ... Li, J. Z. (2020). Persistence and evolution of SARS-CoV-2 in an immunocompromised host. *New England Journal of Medicine*, *383*(23), 2291–2293. https://doi.org/10.1056/nejmc2031364

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., & Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*, *367*(6485), 1444–1448. https://doi.org/10.1126/science.abb2762

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K.-Y., Wang, Q., Zhou, H., Yan, J., & Qi, J. (2020). Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell*, *181*(4), 894–904.e9. https://doi.org/10.1016/j.cell.2020.03.045

Henderson, R., Edwards, R. J., Mansouri, K., Janowska, K., Stalls, V., Gobeil, S. M. C., Kopp, M., Li, D., Parks, R., Hsu, A. L., Borgnia, M. J., Haynes, B. F., & Acharya, P. (2020). Controlling the SARS-CoV-2 spike glycoprotein conformation. *Nature Structural & Molecular Biology*, *27*(10), 925–933. https://doi.org/10.1038/s41594-020-0479-4

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., & Li, F. (2020). Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences*, *117*(21), 11727–11734. https://doi.org/10.1073/pnas.2003138117

Long, Q.-X., Liu, B.-Z., Deng, H.-J., Wu, G.-C., Deng, K., Chen, Y.-K., Liao, P., Qiu, J.-F., Lin, Y., Cai, X.-F., Wang, D.-Q., Hu, Y., Ren, J.-H., Tang, N., Xu, Y.-Y., Yu, L.-H., Mo, Z., Gong, F., Zhang, X.-L., ... Huang, A.-L. (2020). Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nature Medicine*, *26*(6), 845–848. https://doi.org/10.1038/s41591-020-0897-1

Jiang, H.-w., Li, Y., Zhang, H.-n., Wang, W., Yang, X., Qi, H., Li, H., Men, D., Zhou, J., & Tao, S.-c. (2020). SARS-CoV-2 proteome microarray for global profiling of COVID-19 specific IgG and IgM responses. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-17488-8

Parker, E. P. K., Shrotri, M., & Kampmann, B. (2020). Keeping track of the SARS-CoV-2 vaccine pipeline. *Nature Reviews Immunology*, *20*(11), 650–650. https://doi.org/10.1038/s41577-020-00455-1

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., Zhao, C., Zhang, Q., Liu, H., Nie, L., Qin, H., Wang, M., Lu, Q., Li, X., Sun, Q., Liu, J., Zhang, L., Li, X., Huang, W., & Wang, Y. (2020). The impact of mutations in SARS-CoV-2

spike on viral infectivity and antigenicity. *Cell*, *182*(5), 1284–1294.e9. https://doi.org/10.1016/j.cell.2020.07.012

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E. E., Bhattacharya, T., Foley, B., Hastie, K. M., Parker, M. D., Partridge, D. G., Evans, C. M., Freeman, T. M., de Silva, T. I., McDanal, C., Perez, L. G., Tang, H., ... Wyles, M. D. (2020). Tracking changes in SARS-CoV-2 spike: Evidence that d614g increases infectivity of the COVID-19 virus. *Cell*, *182*(4), 812–827.e19. https://doi.org/10.1016/j.cell.2020.06.043

Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., Zhang, X., Muruato, A. E., Zou, J., Fontes-Garfias, C. R., Mirchandani, D., Scharton, D., Bilello, J. P., Ku, Z., An, Z., Kalveram, B., Freiberg, A. N., Menachery, V. D., Xie, X., ... Shi, P.-Y. (2020). Spike mutation d614g alters SARS-CoV-2 fitness. *Nature*. https://doi.org/10.1038/s41586-020-2895-3

Yurkovetskiy, L., Wang, X., Pascal, K. E., Tomkins-Tinch, C., Nyalile, T. P., Wang, Y., Baum, A., Diehl, W. E., Dauphin, A., Carbone, C., Veinotte, K., Egri, S. B., Schaffner, S. F., Lemieux, J. E., Munro, J. B., Rafique, A., Barve, A., Sabeti, P. C., Kyratsous, C. A., ... Luban, J. (2020). Structural and functional analysis of the d614g SARS-CoV-2 spike protein variant. *Cell*, *183*(3), 739–751.e8. https://doi.org/10.1016/j.cell.2020.09.032

Gobeil, S., Janowska, K., McDowell, S., Mansouri, K., Parks, R., Manne, K., Stalls, V., Kopp, M., Henderson, R., Edwards, R. J., et al. (2020). D614g mutation alters sars-cov-2 spike conformational dynamics and protease cleavage susceptibility at the s1/s2 junction. *bioRxiv*.

Rees-Spear, C., Muir, L., Griffith, S., Heaney, J., Aldon, Y., Snitselaar, J., Thomas, P., Graham, C., Seow, J., Lee, N., Rosa, A., Roustan, C., Houlihan, C., Sanders, R., Gupta, R., Cherepanov, P., Stauss, H., Nastouli, E., Doores, K., ... McCoy, L. (2021). The impact of spike mutations on SARS-CoV-2 neutralization. https://doi.org/10.1101/2021.01.15.426849

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N., et al. (2020). Emergence and

rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (sars-cov-2) lineage with multiple spike mutations in south africa. *medRxiv*.

Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D. L., & Volz, E. (2020). *Preliminary genomic characterisation of an emergent sars-cov-2 lineage in the uk defined by a novel set of spike mutations*.

Faria, N. R., Morales Claro, I., Candido, D., Moyses Franco, L., Andrade, P. S., Coletti, T. M., Silva, C. A., Sales, F. C., Manuli, E. R., Aguiar, R. S., et al. (2021). Genomic characterisation of an emergent sars-cov-2 lineage in manaus: Preliminary findings.

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of coot. *Acta Crystallographica Section D - Biological Crystallography*, *66*, 486–501.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The foldx web server: An online force field. *Nucleic acids research*, *33*(suppl_2), W382–W388.

Ray, D., Le, L., & Andricioaei, I. (2021). Distant residues modulate conformational opening in sars-cov-2 spike protein. *bioRxiv*, 2020–12.

Chi, X., Yan, R., Zhang, J., Zhang, G., Zhang, Y., Hao, M., Zhang, Z., Fan, P., Dong, Y., Yang, Y., Chen, Z., Guo, Y., Zhang, J., Li, Y., Song, X., Chen, Y., Xia, L., Fu, L., Hou, L., ... Chen, W. (2020). A neutralizing human antibody binds to the n-terminal domain of the spike protein of SARS-CoV-2. *Science*, *369*(6504), 650–655. https://doi.org/10.1126/science.abc6952

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., & Wang, X. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, *581*(7807), 215–220. https://doi.org/10.1038/s41586-020-2180-5

McCallum, M., Marco, A. D., Lempp, F., Tortorici, M. A., Pinto, D., Walls, A. C., Beltramello, M., Chen, A., Liu, Z., Zatta, F., Zepeda, S., di Iulio, J., Bowen, J. E., Montiel-Ruiz, M., Zhou, J., Rosen, L. E., Bianchi, S., Guarino, B., Fregni, C. S., ... Veesler, D. (2021). N-terminal domain antigenic mapping reveals a

site of vulnerability for SARS-CoV-2. https://doi.org/10.1101/2021.01.14.426475

Greaney, A. J., Loes, A. N., Crawford, K. H., Starr, T. N., Malone, K. D., Chu, H. Y., & Bloom, J. D. (2021). Comprehensive mapping of mutations to the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. https://doi.org/10.1101/2020.12.31.425021

Xie, X., Liu, Y., Liu, J., Zhang, X., Zou, J., Fontes-Garfias, C. R., Xia, H., Swanson, K. A., Cutler, M., Cooper, D., Menachery, V. D., Weaver, S. C., Dormitzer, P. R., & Shi, P.-Y. (2021). Neutralization of SARS-CoV-2 spike 69/70 deletion, e484k and n501y variants by BNT162b2 vaccine-elicited sera. *Nature Medicine*. https://doi.org/10.1038/s41591-021-01270-4

Weissman, D., Alameh, M.-G., de Silva, T., Collini, P., Hornsby, H., Brown, R., LaBranche, C. C., Edwards, R. J., Sutherland, L., Santra, S., Mansouri, K., Gobeil, S., McDanal, C., Pardi, N., Hengartner, N., Lin, P. J., Tam, Y., Shaw, P. A., Lewis, M. G., ... Montefiori, D. C. (2021). D614g spike mutation increases SARS CoV-2 susceptibility to neutralization. *Cell Host & Microbe*, *29*(1), 23–31.e4. https://doi.org/10.1016/j.chom.2020.11.012

Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S., & Crispin, M. (2020). Site-specific glycan analysis of the SARS-CoV-2 spike. *Science*, eabb9983. https://doi.org/10.1126/science.abb9983

Shajahan, A., Supekar, N. T., Gleinich, A. S., & Azadi, P. (2020). Deducing the n- and o-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. *Glycobiology*, *30*(12), 981–988. https://doi.org/10.1093/glycob/cwaa042

Casalino, L., Gaieb, Z., Goldsmith, J. A., Hjorth, C. K., Dommer, A. C., Harbison, A. M., Fogarty, C. A., Barros, E. P., Taylor, B. C., McLellan, J. S., Fadda, E., & Amaro, R. E. (2020). Beyond shielding: The roles of glycans in the SARS-CoV-2 spike protein. *ACS Central Science*, *6*(10), 1722–1734. https://doi.org/10.1021/acscentsci.0c01056

Weisblum, Y., Schmidt, F., Zhang, F., DaSilva, J., Poston, D., Lorenzi, J. C., Muecksch, F., Rutkowska, M., Hoffmann, H.-H., Michailidis, E., Gaebler, C., Agudelo, M., Cho, A., Wang, Z., Gazumyan, A., Cipolla, M., Luchsinger, L., Hillyer, C. D.,

Caskey, M., ... Bieniasz, P. D. (2020). Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife*, *9*. https://doi.org/10.7554/elife.61312

Neher, R. A., Dyrdak, R., Druelle, V., Hodcroft, E. B., & Albert, J. (2020). Potential impact of seasonal forcing on a SARS-CoV-2 pandemic. https://doi.org/10.1101/2020.02.13.20022806

Phelan, J., Deelder, W., Ward, D., Campino, S., Hibberd, M. L., & Clark, T. G. (2020). Controlling the sars-cov-2 outbreak, insights from large scale whole genome sequences generated across the world. *BioRxiv*.

Patiño-Galindo, J. Á., Filip, I., AlQuraishi, M., & Rabadan, R. (2020). Recombination and lineage-specific mutations led to the emergence of sars-cov-2. *bioRxiv*.

Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, *22*(13). https://doi.org/10.2807/1560-7917.es.2017.22.13.30494

Okada, P., Buathong, R., Phuygun, S., Thanadachakul, T., Parnmen, S., Wongboot, W., Waicharoen, S., Wacharapluesadee, S., Uttayamakul, S., Vachiraphan, A., Chittaganpitch, M., Mekha, N., Janejai, N., Iamsirithaworn, S., Lee, R. T., & Maurer-Stroh, S. (2020). Early transmission patterns of coronavirus disease 2019 (COVID-19) in travellers from wuhan to thailand, january 2020. *Eurosurveillance*, *25*(8). https://doi.org/10.2807/1560-7917.es.2020.25.8.2000097

Roberts, E., Eargle, J., Wright, D., & Luthey-Schulten, Z. (2006). Multiseq: Unifying sequence and structure data for evolutionary analysis. *BMC bioinformatics*, *7*(1), 1–11.

Webb, B., & Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Current Protocols in Protein Science*, *86*(1). https://doi.org/10.1002/cpps.20

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004a). UCSF chimera?a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

Hubbard, S. J., & Thornton, J. M. (1993). Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, *2*(1).

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5

Kandt, C., Ash, W. L., & Tieleman, D. P. (2007). Setting up and running molecular dynamics simulations of membrane proteins. *Methods*, *41*(4), 475–488. https://doi.org/10.1016/j.ymeth.2006.08.006

Mark, P., & Nilsson, L. (2001). Structure and dynamics of the TIP3p, SPC, and SPC/e water models at 298 k. *The Journal of Physical Chemistry A*, *105*(43), 9954–9960. https://doi.org/10.1021/jp003020w

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1-2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

Darden, T., York, D., & Pedersen, L. (1993). Particle mesh ewald: AnN·log(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, *98*(12), 10089–10092. https://doi.org/10.1063/1.464397

Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, *126*(1), 014101. https://doi.org/10.1063/1.2408420

Parrinello, M., & Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, *52*(12), 7182–7190. https://doi.org/10.1063/1.328693

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004b). Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*(13), 1605–1612.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., & Ferrin, T. E. (2021). Ucsf chimerax: Structure visualization for researchers, educators, and developers. *Protein Science*, *30*(1), 70–82.

Schrödinger, LLC. (2015). *The PyMOL molecular graphics system, version 1.8* PyMOL The PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC.