Open access • Posted Content • DOI:10.1101/2021.07.16.452571

Molecular evolution and structural analyses of the spike glycoprotein from Brazilian SARS-CoV-2 genomes: the impact of the fixation of selected mutations

Patrícia Aline Gröhs Ferrareze, Ricardo Ariel Zimerman, Vinícius Bonetti Franceschi, Gabriel Dickin Caldana ...+3 more authors

Institutions: Universidade Federal de Ciências da Saúde de Porto Alegre, Universidade Federal do Rio Grande do Sul

Published on: 19 Jul 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Population

Molecular evolution and structural analyses of the spike glycoprotein from Brazilian SARS-CoV-2 genomes: the impact of the fixation of selected mutations

Patrícia Aline Gröhs Ferrareze[1], Ricardo Ariel Zimerman[2], Vinícius Bonetti Franceschi[3], Gabriel Dickin Caldana[1], Paulo Augusto Netz[4], Claudia Elizabeth Thompson[1,3,5]*

[1] Graduate Program in Health Sciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, RS, Brazil

[2] Irmandade Santa Casa de Misericórdia de Porto Alegre, Porto Alegre, RS, Brazil

[3] Center of Biotechnology, Graduate Program in Cell and Molecular Biology (PPGBCM), Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil

[4] Graduate Program in Chemistry, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil

[5] Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, RS, Brazil

*** Corresponding author**

Address for correspondence:

Claudia Elizabeth Thompson

Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), 245/200C Sarmento Leite St, Porto Alegre, RS, Brazil. ZIP code: 90050-170. Phone: +55 (51) 3303 8889.

E-mail: cthompson@ufcspa.edu.br, thompson.ufcspa@gmail.com

Running title: Genomic and structural analyses of spike glycoprotein from brazilian genomes

Keywords: Covid-19, Severe acute respiratory syndrome coronavirus 2, Infectious Diseases, spike, positive selection, molecular evolution

ABSTRACT

The COVID-19 pandemic caused by *Severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) has reached by July 2021 almost 200 million cases and more than 4 million deaths worldwide since its beginning in late 2019, leading to enhanced concern in the scientific community and the general population. One of the most important pieces of this host-pathogen interaction is the spike protein, which binds to the human Angiotensin-converting enzyme 2 (hACE2) cell receptor, mediates the membrane fusion and is the major target of neutralizing antibodies against SARS-CoV-2. The multiple amino acid substitutions observed in this region, specially in the Receptor Binding Domain (RBD), mainly after almost one year of its emergence (late 2020), have enhanced the hACE2 binding affinity and led to several modifications in the mechanisms of SARS-CoV-2 pathogenesis, improving the viral fitness and/or promoting immune evasion, with potential impact in the vaccine development. In this way, the present work aimed to evaluate the effect of positively selected mutations fixed in the Brazilian SARS-CoV-2 lineages and to check for mutational evidence of coevolution. Additionally, we evaluated the impact of selected mutations identified in some of the VOC and VOI lineages (C.37, B.1.1.7, P.1, and P.2) of Brazilian samples on the structural stability of the spike protein, as well as their possible association with more aggressive infection profiles by estimating the binding affinity in the RBD-hACE2 complex. We identified 48 sites under selective pressure in Brazilian spike sequences, 17 of them with the strongest evidence by the HyPhy tests, including VOC related mutation sites 138, 142, 222, 262, 484, 681, and 845, among others. The coevolutionary analysis identified a number of 28 coevolving sites that were found not to be conditionally independent, such as the couple E484K - N501Y from P.1 and B.1.351 lineages. Finally, the molecular dynamics and free energy estimates showed the structural stabilizing effect and the higher impact of E484K for the improvement of the binding affinity between the spike RBD and the hACE2 in P.1 and P.2 lineages, as well as the stabilizing and destabilizing effects for the positively selected sites.

INTRODUCTION

The *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) is the etiological agent of the COVID-19 pandemic. The virus is composed of an enveloped positive-sense single-stranded RNA genome, which encodes 16 non-structural proteins (nsp1-16), four structural proteins (Spike, Membrane, Envelope, and Nucleocapsid), and other accessory proteins (ORFs 3a, 6, 7a, 7b, 8, 10) (Fehr and Perlman, 2015; Liu et al., 2014). The spike (S) glycoprotein is necessary for the viral binding to ACE2 host cell receptor, it is common to multiple coronaviruses (*e. g.,* MERS-CoV and SARS-CoV) (Fehr and Perlman, 2015), and is the major target of neutralizing antibodies against SARS-CoV-2 (Walls et al., 2020; Yuan et al., 2020). Spike has a homotrimeric form where each monomer (protomer) is composed by a S1 subunit that mediates the receptor binding and an alpha-helix rich S2 subunit that promotes the subsequent membrane fusion (Hoffmann et al., 2020).

Since the emergence of the first SARS-CoV-2 in the early Wuhan epidemic, several mutations have been identified, both occurring isolated or as a signature of a broader mutational complex. Although S protein represents only 13% of the viral genome, substitution at this site has been overrepresented suggesting immunodominance of this protein, specially at its Receptor Binding Domain (RBD) (Greaney et al., 2020; Weisblum et al., 2020). These mutations can enhance receptor binding, either directly or allosterically, alter viral fitness or promote immune evasion, ensuing occasional reinfection and possible impacting the efficacy of developed vaccines (Greaney et al., 2020; Harvey et al., 2021; Korber et al., 2020; Weisblum et al., 2020).

The S protein mutation D614G (aspartic acid to glycine substitution at amino acid position 614) has been progressively dominant worldwide since March 2020, playing an essential role in infectivity and augmented viral load (Groves et al., 2021; C. B. Jackson et al., 2021; Korber et al., 2020; Yurkovetskiy et al., 2020). This substitution may have emerged independently and promptly became dominant in several viral strains, outcompeting D614

harboring viruses, even where they originally appeared. D614G abolishes a hydrogen bond between this position and a threonine present in S2 of the neighbour protomer. The consequent conformational changes acts allosterically favouring the maintenance of the Receptor Binding Domain (RBD) in an activated “up” or “open” position, which exhibits its ACE2 binding site (Receptor Binding Motif) for longer periods. This allows a more efficient binding to ACE2 (Groves et al., 2021), with consequent higher replication and viral loads. However, G614 mutants are similarly (or even more) susceptible to immune neutralization as the original D614 variant (Plante et al., 2020; Weissman et al., 2021). Notably, this enhanced epitope exposure of “up” RBD may be a “weakness”, since RBD is poor in predicted O and N-linked glycan sites required for immune shielding. In theory, this undesirable effect could be compensated by accumulation of other RBD mutations, leading to increased infectiousness, and rendering D614G unnecessary. Specifically, the “HMN 19B variant”, recently described in France, is a direct descendent from the earlier 19B that have predominated before D614G harboring lineages took over (Fourati et al., 2021).

The codon 501 in RBD has been considered another major mutational hotspot. Substitutions at this position have been associated with increased binding affinity to ACE2 (Gu et al., 2020; Starr et al., 2020). The non-synonymous mutation from an asparagine to a tyrosine at position 501 (N501Y) has first appeared in samples from Wales and other parts of the world. However, its actual emergence and dissemination have been related to coevolution with different mutations. The deletion of H69-V70 in the N-terminal-domain, co-occurring in B.1.1.7 lineage, has been particularly important for the N501Y emergence (Meng et al., 2021). However, N501Y has increasingly been demonstrated in coevolution with many different mutational signatures, for instance, in the Variants of Concern (VOCs) B.1.351 and P.1, as well as in the more recent HMN 19B aforementioned. Data suggests that some N501Y harboring viruses could be 50% more transmissible and up to 61% more lethal, mainly due to major changes in the electrostatic interaction between Spike and ACE2 receptors (Washington et al.,

2021; Davies et al., 2021). This would be related to the substitution of a relatively small asparagine (N) for a larger aromatic tyrosine (Y), allowing for an extra contact site with ACE2 (Ali et al., 2021; Gu et al., 2020; Tang et al., 2021). Interesantly, N501Y increases host diversity allowing direct infection of non-humanized mice (Rathnasinghe et al., 2021).

The E484K mutation in RBD has been drowning progressive attention. It is an amino acid replacement of glutamic acid to lysine at amino acid position 484, which could significantly alter the complementarity of antibodies to the RBD region leading to immune evasion (Baum et al., 2020; Greaney et al., 2020; Weisblum et al., 2020). In fact, both in a Deep Mutational Scanning (DMS) experiment and in a selective pressure study using 19 types of mAbs, different substitutions at this position (E484K/Q/P/A/D/G) were consistently demonstrated as being the most critical for decreasing antibody neutralization titers (Greaney et al., 2020; Harvey et al., 2021). Since this mutation has been arising independently in multiple lineages (Ferrareze et al., 2021), at the beginning it may represent a common evolutionary solution for viral maintenance. E484K emerged in a few selected lineages where it co-exists with other signature substitutions (e.g. in B.1.351 and in P1), similarly to the phenomenon previously described for N501Y. Importantly, the combination of E484K and N501Y seems to induce more conformational changes than the N501Y mutant alone, potentially altering antibodies binding to this region and resulting in the immune evasion phenomena (Nelson et al., 2021).

The emergence of three independently evolving lineages (B.1.1.7, B.1.351, and P.1) (Faria et al., 2021; Rambaut et al., 2020; Tegally et al., 2021), all characterized by a constellation of mutations (including the aforementioned del 69-70, K417N/T, E484K, and N501Y convergent mutations) in RBD, has prompted concerns about the evolutionary forces leading to the SARS-CoV-2 adaptation. The presence of these sets of consistently present substitutions are highly suggestive of coevolutionary mutational processes (Martin et al., 2021). In addition to these convergent mutations, other substitutions of these VOCs are also associated with positive selection. For example, there is evidence that eight out of the 10

lineage-defining mutations in the spike protein of the P.1 lineage are under diversifying positive selection (Faria et al., 2021). It is possible to detect selection along prespecified lineages that affect certain subsets of codons in a protein-coding gene. Sites 484 and 501 (located in the RBD) definitively show a pattern of nucleotide diversification that can be linked with positive selection (Tegally et al., 2021). On the other hand, alternative evolutionary processes that could have led to this plethora of substitutions have been rarely described thus far. Recombination, a recurrent and well documented process of the molecular evolution of coronaviruses, has been described between B.1.1.7 and other lineages (Jackson et al., 2021), but their real role in SARS-CoV-2 evolution remains elusive.

The present work aimed to evaluate the effect of positively selected mutations fixed in the Brazilian SARS-CoV-2 lineages and to check for mutational coevolution evidence. Additionally, we evaluated the impact of some mutations identified in some VOC and VOI lineages (C.37, B.1.1.7, P.1 and P.2) of Brazilian samples on the structural stability of the spike protein, as well as their possible association with more aggressive infection profiles by the binding affinity estimation in the RBD- hACE2 complex.

METHODOLOGY

SPIKE SEQUENCE ANALYSES

Genome sequence retrieving

After the exclusion of incomplete or low coverage (Ns > 5%) sequences, 11,078, Brazilian genomes were recovered from the GISAID database between January 01st, 2020 and June 06th, 2021 (submission date up to June 06th, 2021).

Spike phylogenetic analysis

The multiple sequence alignment with NC_045512.2 as reference was performed by the MAFFT v.7 web server (Katoh et al., 2019) with default parameters and 1PAM / κ=2' scoring

matrix for closely related DNA sequences. For the spike sequence analysis, the region between the positions 21,562 and 25,384 (spike regions in the NC_045512.2 reference genome) were selected from the multiple sequence alignment previously performed. The deletion of sequence duplicates (identical spike sequences) kept 2,901 unique spike sequences. The clade assignment and variant calling was performed by Nextclade (https://clades.nextstrain.org/). The phylogenetic analysis was started by the inference of the best evolutionary model by ModelTest-NG (Darriba et al., 2020), which identified GTR+R4 in all selection strategies. The phylogenetic tree reconstruction was performed by the Maximum Likelihood method in the IQ-TREE program (Nguyen et al., 2015), using 1,000 replicates of ultrafast bootstrap (Hoang et al., 2018) and a Shimodaira-Hasegawa-like approximate likelihood ratio test (SH-aLRT) with 1,000 replicates (Guindon et al., 2010), 2,000 iterations and the optimization of the UFBoot trees by NNI on bootstrap alignment. The tree visualization was performed by FigTree software (http://tree.bio.ed.ac.uk/software/figtree/).

Spike selection and coevolutionary analyses

The multiple sequence alignment of the 2,901 unique spike sequences and the phylogenetic tree previously built were used as input in the HyPhy program. To perform different selection tests, the methods: (i) Fast Unconstrained Bayesian AppRoximation (FUBAR) (Murrell et al., 2013), (ii) Fixed Effects Likelihood (FEL) (Kosakovsky Pond and Frost, 2005), and (iii) Single-Likelihood Ancestor Counting (SLAC) (Kosakovsky Pond and Frost, 2005) were evaluated. The analysis of coevolution across sites in the spike sequences was performed by BGM (Bayesian Graphical Model) (Poon et al., 2007), a tool for detecting coevolutionary interactions between amino acid positions in a protein by a Markov Chain Monte Carlo (MCMC) method.

SPIKE STRUCTURAL ANALYSES

Spike structural stability

The analysis of the spike structural stability by changes in free energies upon mutation (ΔΔG kcal/mol) was performed using PDB ID 6XR8 (chain A) as the structure of the full-length prefusion conformation of the SARS-CoV-2 spike protein (Cai et al., 2020) using five methodologies: the web server DynaMut (Rodrigues et al., 2018), FoldX Suite v5.0 (Schymkowitz et al., 2005), and the web servers iMutant3.0 (Capriotti et al., 2005), MAESTRO (Laimer et al., 2016), and PremPS (Chen et al., 2020).

DynaMut: The web-server DynaMut was selected to perform the analyses of vibrational entropy and total energy using the PDB ID 6XR8 (chain A). The DynaMut implements a Normal Mode Analysis (NMA) through Bio3D (Grant et al., 2006) and ENCoM (Frappier and Najmanovich, 2014) approaches, providing rapid and simplified access to insightful analyses about protein motions. Moreover, DynaMut also enables rapid analysis of the impact of mutations on dynamics and stability of proteins resulting from vibrational entropy changes.

FoldX: For the energy estimation, the analysis and correction of structural problems resulting from crystallography in PDB ID 6XR8 were performed using the RepairPDB command and additional parameters *--water=crystal*, *--pH=7.4* and *--temperature=309.65*. Posteriorly, the mutagenesis process was applied for all the positively selected sites with the BuildModel function (additional parameters: *--water=crystal*, *--pH=7.4*, *--temperature=309.65* and *--numberOfRuns=5*). The average values for each ΔΔG (based on the 5 runs) were classified in seven categories according to the reported Foldx accuracy (0.46 kcal/mol): highly stabilizing mutations (ΔΔG < −1.84 kcal/mol), stabilizing mutations (−1.84 kcal/mol ≤ ΔΔG < −0.92 kcal/mol), slightly stabilizing mutations (−0.92 kcal/mol ≤ ΔΔG < −0.46 kcal/mol), neutral mutations (−0.46 kcal/mol < ΔΔG ≤ +0.46 kcal/mol), slightly destabilizing mutations (+0.46 kcal/mol < ΔΔG ≤ +0.92 kcal/mol), destabilizing mutations (+0.92 kcal/mol < ΔΔG ≤ +1.84 kcal/mol), and highly destabilizing mutations (ΔΔG > +1.84 kcal/mol) (Studer et al., 2014).

iMutant3: The iMutant web server was selected to estimate the free energy changes by a support vector machine (SVM)-based tool. This system predicted the sign of the protein stability change upon mutation and as a regression estimator predicted the related ΔΔG values in the physiological pH (7.4) and temperature (36.5 ℃).

MAESTRO: The Multi AgEnt STability pRedictiOn web server was used to estimate the changes in unfolding free energy upon point mutation through a machine learning system.

PremPS: The estimation of the unfolding Gibbs free energies was performed by a random forest regression scoring function using the ProTherm database for parameterization. Negative values indicate stabilization by the decrease of the free energies.

Spike RBD comparative homology modelling

To perform an accurate estimation of the binding free energy associated with mutations in the spike protein belonging to different lineages, spike RBD - ACE2 protein complexes were modelled for the reference SARS-CoV-2 spike (YP_009724390.1) and lineages C.37, B.1.1.7, P.1, P.2, and P.2+452. The PDB file 6M0J was selected as the best template (2.45 Å and 194 amino acids) to the modeling using the MODELLER pipeline (Webb and Sali, 2016). Five models were generated for each sequence. The thirty resulting structures were evaluated in relation to stereochemical parameters and structural quality by the programs PROCHECK (Laskowski et al., 1993) and VERIFY3D (Eisenberg et al., 1997), available on SAVES v6.0 web server (https://saves.mbi.ucla.edu/). The analysis of the Ramachandran plot statistics, G-factors and residue properties (PROCHECK), as well as the evaluation of the 3D-1D score using VERIFY3D, allowed the selection of the best modelled RBD structures to be subsequently used for the energy calculations. The generation of the spike RBD - ACE2 complexes for each different lineage was performed by the PyMOL software using the structural alignment of the RBD models with the 6M0J template, which was the source of the ACE2 structural coordinates.

Molecular dynamics and binding free energy estimation

The structures of the fragments comprising residues 333 until 526 of the wild-type (reference) spike protein, as well as of the variants P.1 (K417T, E484K, N501Y), P.2 (E484K), P.2+452 (E484K, L452V), C.37 (L452Q, F490S), and B.1.1.7 (N501Y) complexed with the human ACE2 protein (residues 19 until 615) in the pdb format were used as input for classical molecular dynamics simulations using GROMACS (Abraham et al., 2015) with the AMBER03 force field (Duan et al., 2003).

The spike-ACE2 complexes were simulated in cubic boxes with periodic boundary conditions, solvated with TIP3P water molecules (Jorgensen et al., 1983) and with sodium and chloride ions corresponding to physiological concentration. The van der Waals interactions were calculated using a cutoff radius of 1.2 nm and the electrostatic interactions were calculated with the Particle Mesh Ewald (PME) method (Darden et al., 1993). All systems were initially energetically minimized using conjugate gradients and steepest descent algorithms. After minimization, they were submitted to a 500 ps where the coordinates of both proteins were restrained, allowing the solvent and ions to relax without disturbing the geometry of the complex. After the position restrained simulation, a thermalization phase consisting of a sequence of three unrestricted molecular dynamics simulations with 5 ns each in temperatures of 200 K, 240 K, and 280 K was carried out. The production phase consisted of 200 ns long simulation runs, in the NPT ensemble, using a Nosé-Hoover thermostat (Hoover, 1985; Nosé, 1984) and a Parrinello-Rahman barostat (Parrinello and Rahman, 1981).

After simulations, the trajectories were visually analyzed using VMD (Humphrey et al., 1996) and GROMACS tools to quantify the structural and thermodynamic stability of the complexes, the number of hydrogen bonds and the interface area. The interface area was computed taking SASA (Solvent Accessible Surface Area) (Eisenhaber et al., 1995) of ACE2 and spike minus SASA of the complex and dividing the result by two (because the contact was counted twice).

Spike RBD-ACE2 binding free energy calculation

The spike-ACE2 binding free energy was calculated using Molecular Mechanics Poisson-Boltzmann Surface Area (MM/PBSA) binding free energy calculations (Baker et al., 2001; Homeyer and Gohlke, 2012), employing sets of 200 configurations (1ns spaced snapshots obtained from the molecular dynamics trajectories) for each system. The calculations were carried out using the program g_mmpbsa (Kumari et al., 2014), which is compatible with GROMACS, using as settings a gridspace of 0.5 A, salt concentration of 0.150 M, solute dielectric constant of 2, and estimating the nonpolar solvation energy using the solvent accessible surface area (SASA). The contributions of the residues to the binding free energy were calculated with the same program.

RESULTS

Spike protein phylogenetic analyses

The spike phylogenetic analysis of 2,901 unique sequences showed the formation of a large monophyletic group containing two main subclades: one formed by P.1 and P.1.2 sequences and another formed by the B.1.1.7. B.1.1.28 and P.2 sequences were located at the basis of the P.1 subclade, while some B.1.1.33/N.9 sequences are basal to the B.1.1.7 subclade. Interesting to note that B.1.1.7 formed one monophyletic group (Figure 1A). Based on spike sequences, the P.2, B.1.1.28 and B.1.1.33 lineages did not form individual clades. Finally, it was possible to observe the formation of a large basal clade with B.1.1.28 and B.1.1.33 sequences, with the B lineage (among others) as its ancestor. Despite P.2 being considered as derivative from B.1.1.28 by the inclusion of the E484K mutation (D614G+E484K+V1176F), the presence of this substitution along the tree (as well D614G and V1176F) indicated that the formation of all the small clades is probably due to the high variable combinations of mutations found in these lineages.

The genetic analysis of these 2,901 unique Brazilian spike sequences showed that 575 amino acid sites presented missense substitutions in different lineages, with a mutation rate of 7.79 amino acid substitutions per spike sequence/genome, since January 2020. With a range between 0 and 16 amino acid substitutions and a different average count in each lineage (Figure 1B), the spike sequence set was mostly represented by lineages such as P.1 (45.71%), P.2 (14.34%), B.1.1.28 (13.82%), B.1.1.33 (10.62%), and B.1,1.7 (4.03%), from 53 identified lineages. The evaluation of the mutation rate before and after the P.1 first occurrence (October 2020) indicated that the spike sequences showed an estimated amino acid substitution of 2.12 events per genome between January and September, 2020. However, from October 2020 up to June 06 2021, this rate was increased to 8.99 amino acid mutations for each spike sequence.

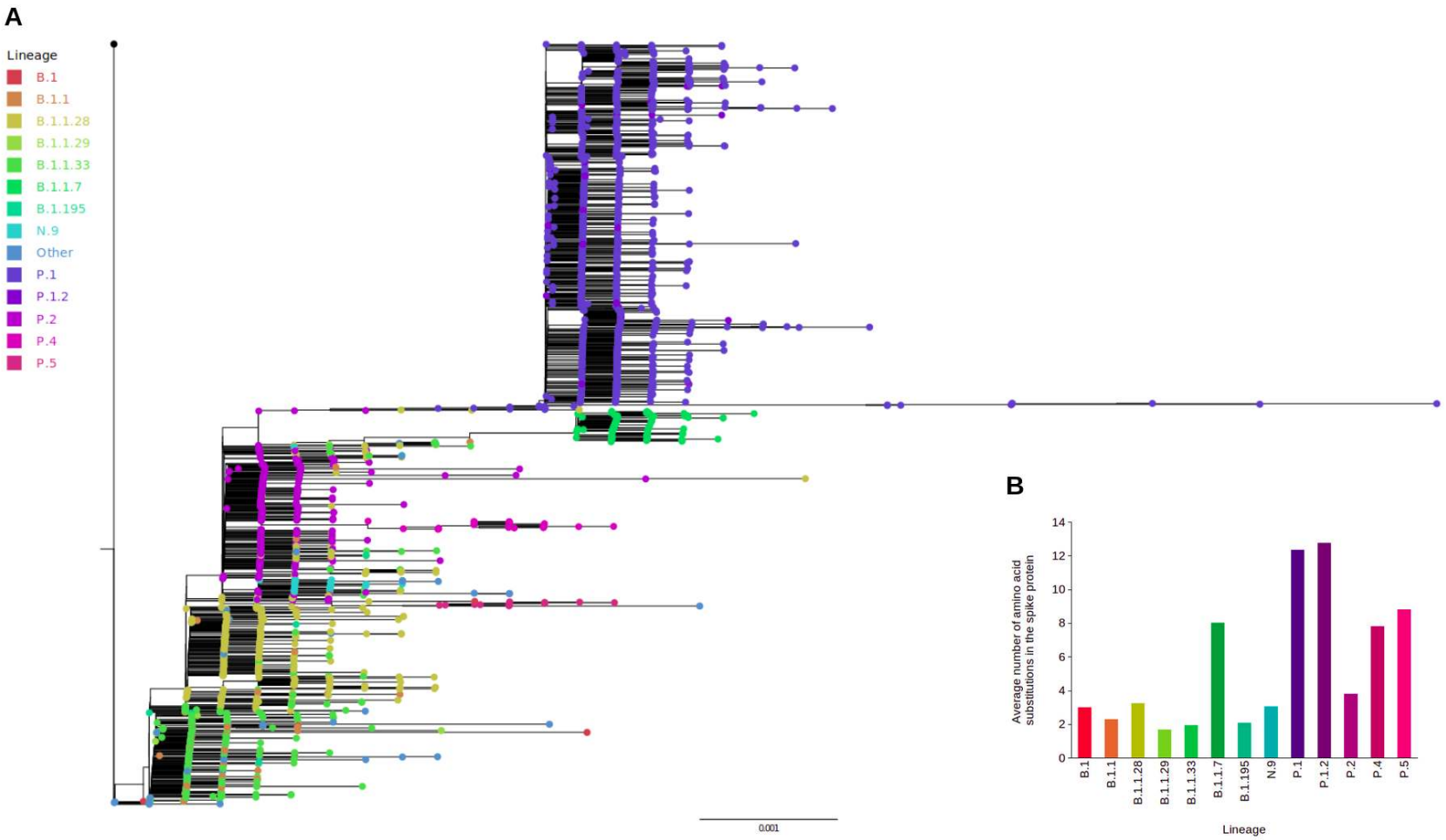

Figure 1. Spike phylogenetic and genomic analyses. (A) Phylogenetic tree of 2,901 unique Brazilian spike nucleotide sequences available until June 06, 2021. Node tips are colored by PANGO lineages represented by ≥ 10 genomes. “Other” defines lineages representing < 10 genomes. The tree is rooted using the reference spike nucleotide sequence (NC_045512.2). (B) Average number of amino acid substitution events in the spike protein by lineage. The average values were calculated based on the spike set of 2,901 brazilian sequences between January, 2020 and June 06, 2021. This spike set represents the nucleotide sequence variability for 11,054 brazilian genomes.

Spike selection analyses

The selection analysis performed with HyPhy for site-to-site tests evaluated the presence of diversifying and purifying selection in the spike protein using a set of 2,901 sequences, which represents the genetic variability for 11,054 genomes (24 sequences were excluded due to the presence of >5% of ambiguous sites or truncating insertions in the final alignment). The FUBAR method identified 20 sites under adaptive pressure along the phylogeny (Table 1 and Supplementary File 2). Among these, the mutation in site 484 is represented by three missense substitutions: (i) E484K, which was found in 71.36% of all genomes, belonging to 11 different lineages, (ii) E484Q, present in 0.22% of the sequences, including the B.1, B.1.1.28 and P.5 lineages, and (iii) a new substitution found in only one genome, E484D. Other sites related to known mutations from VOC lineages such as P.1, B.1.1.7, B.1.351, and B.1.617.2 were identified as potentially fixed substitutions. In this group, there are sites with the mutations L5F (0.95% of the spike sequences), A67S/V/P (0.12/0.08/0.01%), G75V (0.11%), D138Y/H (55.04/0.11%), G142D (0.03%), A222V/S/P (0.49/0.05/0.01%), A262S/D (0.50/0.04%), K417T/N (49.11/0.05%), N501Y (61.12%), T572I (0.34%), P681H/L/R (3.03/0.40/0.03%), A688V/S/G (0.19/0.04/0.03%), A845S/V (1.27/0.04%), V1176F (82.12%), among others.

The FEL method was mainly tested to detect negatively selected sites, since more powerful methods such as FUBAR do not evaluate purifying selection. With the initial assumption that the selective pressure for each site is constant along the entire phylogeny, FEL indicated sites that are evolving under positive and negative selective pressures in the spike protein sequence. As result, 239 sites were marked as targets of purifying selection (Supplementary File 3) along the phylogeny, while other 44 sites were suggested to be under adaptive selection. Interestingly, all positively selected sites indicated by the FUBAR method were found by FEL (Table 1 and Figure 2A), including those with low frequency in the analysed genomes. Some of them were indeed related to the VOC lineages. Among the sites identified

only by the FEL test are the residues 29, 49, 63, 76, 78, 367, 483, 553, 585, 689, 747, 1,027, 1,084, 1,124, 1,133, and 1,167 (Supplementary File 2).

Finally, using a modified version of the Suzuki-Gojobori counting algorithm, the SLAC method assumes that the selective pressure for each site is constant. Despite being the weakest method of this group (Kosakovsky Pond and Frost, 2005), the analysis of pervasive site-selection with SLAC indicated the presence of 130 sites under negative selection, of which 112 were previously found by FEL (Supplementary File 3 and Figure 2B). In relation to the 29 positively selected sites, 17 of them were previously identified by the FUBAR and FEL methods (Table 1), with eight also found by FEL (12, 26, 98, 570, 653, 684, 846, and 1,078) and four only identified by SLAC (27, 701, 769, and 1,228). Considering the consensus sites identified by the three selection methods, the residues 5, 21, 67, 75, 138, 142, 222, 262, 484, 572, 614, 681, 688, 845, 1,176, 1,219, and 1,264 presented the strongest evidences for positive selection pressure on the spike protein (Table 1 and Figure 2A). Following are those recognized by two methods such as sites 12, 26, 98, 417, 501, 570, 653, 684, 1,078, and 1,260.

These missense substitutions were identified in up to 15 different lineages (except for mutations in site 614). This is the case of V1176F, which arose as a B.1.1.28 lineage-defining mutation and is now spread along the phylogenetic tree. The arithmetical average calculation (without site 614) showed an occurrence of 3.22 lineages per each amino acid substitution type, while N501Y was found in 10 lineages, N501T was described in only four. Despite the occurrence of different substitutions in the same site for different lineages, generating an average of 2.19 amino acid mutation possibilities for each mutated site, some lineages were predominantly found. From the 37 identified lineages (excluding those from site 614), the variability of P.1 genomes comprised 45 of the 48 sites under adaptive selection considering all predicted sites by FUBAR, FEL, and SLAC, in some cases, covering multiple substitutions on the same site (e.g.: A67S/V/P, P681H/L/R). For the sites predicted by the three methods (n =

17), the lineages P.2 and B.1.1.28 were identified in the analysis of all events (Supplementary File 2).

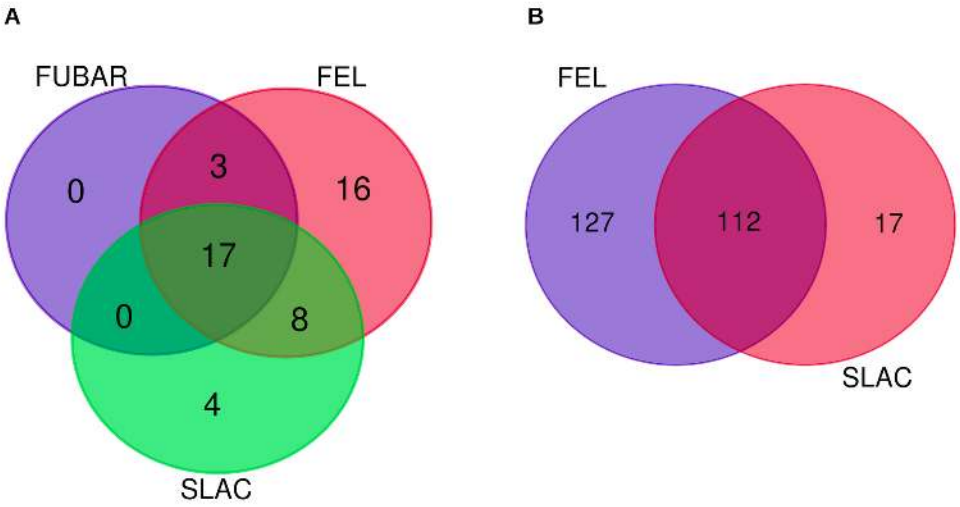

Figure 2. Number of predicted sites under adaptive and purifying selection by each HyPhy method. (A) Positively selected sites shared by FUBAR, FEL, and SLAC methods. (B) Sites under purifying selection detected by FEL and SLAC.

Table 1. Positively selected sites detected by all tested HyPhy evolutionary methods (FUBAR, FEL and SLAC) for pervasive site-level selection and its frequencies (%) in the genome set (n=11,078).

Nr. Lineages	Mutation (%)	Site	FUBAR			FEL				SLAC		
			α	β	post. p	α	β	LRT	p	dS	dN	p
7	L→F 0.95	5	0.465	15.525	1.0000	0.000	11.735	19.399	0.0000	0.000	17.956	0.000
4	R→I 0.08	21	1.150	6.195	0.9034	0.000	4.701	2.988	0.0839	0.000	4.998	0.030
2	R→T 0.02											
1	R→S											

	0.01											
6	A→V 0.12	67	0.655	6.606	0.9824	0.000	4.581	7.605	0.0058	0.000	6.500	0.005
5	A→S 0.08											
1	A→P 0.01											
5	G→V 0.11	75	0.667	6.269	0.9727	0.000	4.404	7.260	0.0070	0.000	5.999	0.008
2	G→S 0.02											
1	G→D 0.01											
6	D→Y 55.04	138	0.523	50.000	1.0000	0.000	38.079	38.743	0.0000	0.000	35.868	0.000
4	D→H 0.11											
4	G→V 0.77	142	0.734	4.497	0.9208	0.000	3.684	6.030	0.0141	0.000	5.498	0.012
2	G→D 0.03											
2	G→S 0.02											
8	A→V 0.49	222	1.394	7.103	0.9682	1.043	5.642	4.467	0.0345	1.000	8.000	0.010
1	A→S 0.05											
1	A→P 0.01											
3	A→S 0.50	262	0.634	7.286	0.9974	0.000	6.372	10.494	0.0012	0.000	9.001	0.001
4	A→D 0.04											
3	A→G 0.04											
11	E→K 71.36	484	0.842	49.941	0.9999	0.000	35.319	11.838	0.0006	0.000	23.427	0.000
3	E→Q 0.22											
1	E→D 0.01											
5	T→I 0.34	572	0.743	4.142	0.9115	0.000	3.615	5.771	0.0163	0.000	3.996	0.039

2	T→N 0.02											
56	D→G 99.14	614	0.698	5.468	0.9493	0.000	4.044	6.636	0.0100	0.000	5.984	0.008
1	D→S 0.01											
8	P→H 3.03	681	0.659	6.650	0.9833	0.000	4.507	7.461	0.0063	0.000	6.936	0.004
3	P→L 0.4%											
2	P→R 0.03											
2	A→S 0.04	688	0.745	4.091	0.9100	0.000	3.519	5.829	0.0158	0.000	5.000	0.017
6	A→V 0.19											
1	A→G 0.03											
4	A→S 1.27	845	0.746	4.085	0.9098	0.000	3.512	5.825	0.0158	0.000	5.000	0.017
1	A→V 0.04											
15	V→F 82.12	1176	0.522	26.249	1.0000	0.000	16.344	19.637	0.0000	0.000	21.365	0.000
6	G→C 0.28	1219	1.334	8.108	0.9758	1.047	7.341	6.702	0.0096	1.004	9.985	0.002
4	G→V 0.11											
5	V→L 0.30	1264	0.900	6.706	0.9788	0.586	4.766	7.450	0.0063	1.001	6.498	0.028
1	V→M 0.01											

FUBAR inferred 20 sites submitted to diversifying selection at posterior probability of ≥ 0.9. Of these, 0.69 are expected to be false positive (95% confidence interval of 0-3). FEL found 44 sites under pervasive positive and 239 sites under negative selection at $p \leq 0.1$. SLAC found 29 sites under pervasive positive and 130 sites under negative selection at $p \leq 0.1$.

Spike coevolutionary analysis

In order to evaluate if the spike protein sites could be under a coevolutionary process, we performed a HyPhy BGM analysis. The Bayesian Graphical Model (BGM) method is a tool

for detecting coevolutionary interactions between amino acid positions in a protein. This method works by detecting pairs of positions with correlated mutations in protein multiple sequence alignments. It also performs a statistical analysis of the distribution of mutations in the branches of the tree. The coevolutionary analysis detected mutational correlations between 62 pairs of sites (Supplementary File 4), of which 28 achieved a P ≥ 0.8 (Table 2). Of these, sites 262, 484, 501, and 681 were identified under diversifying selection and are suggested to coevolve in a conditional manner. Specifically, sites 484 and 501 seem to coevolve. In fact, the E484K + N501Y combination is found in three VOCs: B.1.351, P.1, P.1.1 and P.1.2, besides others such as the Variants of Interest (VOIs) as P.2. For the B.1.1.7 lineage, although the E484K mutation is not lineage-defining, it can be present (with N501Y) in some variants.

Table 2. Coevolutionary analysis: Bayesian Graphical Model (BGM) inference to substitution histories at individual sites with P ≥ 0.8.

Site 1	Site 2	P [Site 1 ↔ Site 2]	Subs (1, 2, shared)
1	1130	0.848	3, 1, 1
9	445	0.815	3, 1, 1
88	603	0.908	2, 1, 1
90	91	0.801	4, 1, 1
92	93	0.966	1, 1, 1
147	1070	0.819	3, 1, 1
155	670	0.918	1, 2, 1
170	171	0.921	2, 1, 1
180	764	0.815	4, 1, 1
181	849	0.839	4, 1, 1
188	673	0.868	3, 1, 1
210	445	0.805	3, 1, 1
226	228	0.919	1, 2, 1
232	233	0.966	2, 2, 2
244	1165	0.801	1, 4, 1

262	265	0.987	18, 2, 2
265	266	0.932	2, 1, 1
312	314	0.996	2, 5, 2
330	1042	0.922	2, 1, 1
343	344	0.915	1, 2, 1
484	**501**	0.893	51, 8, 3
528	938	0.966	1, 1, 1
549	575	0.804	1, 4, 1
574	637	0.826	4, 1, 1
630	632	0.874	1, 3, 1
639	946	0.801	2, 2, 1
681	716	0.962	14, 4, 2
777	779	0.845	1, 2, 1

BGM analysis summary on 574 sites with at least one substitution. Evidence for conditional dependence was reported at posterior probability of 0.5: 62 pairs of conditionally dependent sites identified. The positively selected sites are indicated in bold. P [Site 1 ↔ Site 2]: Probability that the sites 1 and 2 are not conditionally independent. Subs (1, 2, shared): Substitution counts inferred for site 1, 2 and substitutions shared by both sites.

Spike structural analyses

To analyze the impact of the spike sites under adaptive selection on the protein stability and host-pathogen interaction, two main categories of structural analyses were applied related to: (i) the folding/unfolding free energies and the vibrational entropy: a reference spike structure PDB ID 6XR8 (relative to the NC045512.2 genome) was used to estimate the energy changes caused by these positively selected single mutations in the prefusion conformation of the spike protein and (ii) protein molecular dynamics: SARS-CoV-2 spike RBD models belonging to different viral lineages in a complex with the human ACE2 were simulated using molecular dynamics and submitted to estimation of binding free energies to provide a better understanding of the effect of those sites under positive selection (single and combined) in the viral fitness.

Spike structural stability

The evaluation of the structural stability of the spike protein considered the results of five different methodologies in order to find a majority consensus. The application of DynaMut, FoldX, iMutant, MAESTRO, and PremPS provided the estimate of the unfolding and total free energies, as well as the vibrational entropy of the mutated structures. Despite some sites presenting opposite trends in relation to the stabilizing/destabilizing impact, which denotes the specific differences in energy calculations among these algorithms (Figure 3A), a majority consensus result (found by 3 or more tests) was obtained (Figure 3B and Table 3). Sites without resolution in the crystallographic structure, located at N/C-terminal regions, as those in the furin cleavage site (among others), were not analyzed for the energy estimation.

The PremPS algorithm was used to calculate the changes in unfolding Gibbs free energy generated by each single mutation in the full-lenght prefusion conformation of the spike protein. As observed in Table 3, important mutations for the VOC lineages B.1.1.351, B.1.1.7 and P.1, such as E484K and N501Y, seem to stabilize the protein structure by the decrease of the unfolding free energy change (-0.16 and -0.89 kcal/mol, respectively). Substitutions as those from residues 138 and 585 presented the lowest ΔΔG values, which may suggest a higher stabilizing effect of the mutations D138Y and L585F in lineages as P.1 and P.2, among others. However, for site 585, a destabilizing effect was found by three methods (iMutant, DynaMut and FoldX).

The analysis of the impact of the positively selected mutations on protein dynamics and stability was performed with the DynaMut web-server. The higher vibrational entropy ($\Delta\Delta G_{Vib}$ ENCoM) found in sites 21 (R→S), 78 (R→S/T) and 417 (K→T/N) indicated an increased molecular flexibility (>0.5 kcal/mol) due to the loss of molecular interactions. Moreover, the consensus results correlated with the predicted destabilizing impact. In other sites, such as 49, the conformational molecular rigidity caused by a reduced vibrational entropy (-2.809 kcal/mol) follows the stabilizing effect predicted by DynaMut, MAESTRO, and FoldX. Especially, site 689

achieved the lowest $\Delta\Delta G_{vib}$, with an estimated reduction in the molecular flexibility of -4.514 kcal/mol in the serine to isoleucine substitution. Although the substitution of a negatively charged glutamic acid by a positively charged lysine increased the vibrational entropy in the E484K-mutated protein, slightly decreasing the molecular rigidity ($\Delta\Delta G_{vib}$ = +0.246 kcal/mol), a potential stabilizing effect of this mutation was predicted by PremPs, Dynamut, MAESTRO, and FoldX analyses.

The iMutant3.0 web server was used to evaluate the changes in the thermodynamic stability of folded proteins. Using the physiological pH (7.4) and temperature (36.5 ℃) as parameters, the SVM algorithm provided the estimation of the free energy change ($\Delta\Delta G$) between the wild-type and the mutated 6XR8 spike structures. According to the analysis of the thermodynamic properties, the mutations marked as positively selected seem to destabilize the protein structure. However, the majority consensus evaluation did not confirm this trend, suggesting 51 substitution events (from 33 sites) as stabilizing mutations.

Finally, the FoldX pipeline repaired the 6XR8 PDB file by the correction of residues with bad torsion angles or van der Waals clashes and performed the energy minimization testing different rotamer combinations. As result of the mutagenesis and total energy estimation, the mutations in sites T29A, G142D, and A570D were considered as highly destabilizing, increasing the total energy in 1.936, 2.848 and 2.939 kcal/mol, respectively. Despite some energy differences being categorized as neutral, the increase or decrease of the math signal suggests the potential effect of these mutations on the structural stability of spike, in the tested conditions.

The comparison between the mutational effect predicted by the five methods and the majority consensus result are presented in Figure 4. The evaluation of all existing substitutions for each positively selected site by the majority consensus indicated the prevalence of stabilizing mutations, which was observed for the important VOC and VOI mutated sites. The analysis of the positively selected sites according to the prevalent substitution per site obtained

a similar profile, with 25.71% of the mutated sites generating a destabilizing effect in the spike prefusion structure (Figure 5).

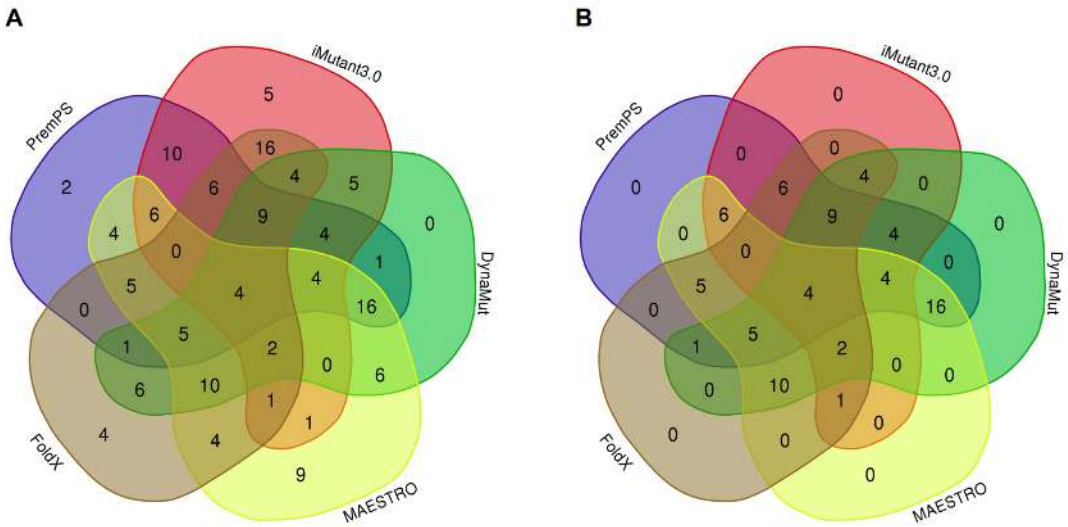

Figure 3. Results for the effect of mutations on the spike protein using different methods. **(A)** Distribution of the number of mutations and their stabilizing/destabilizing effect for each tested method considering 77 mutations in 35 positively selected sites. **(B)** Analysis of all 77 mutations in 35 positively selected sites, considering only the majority consensus, i.e. stabilizing/destabilizing effect appearing in three or more methods.

Table 3. Energy estimation for positively selected mutations.

			PremPS Unfolding Free Energy	**iMutant3.0 Folding Free Energy**	**DynaMut Vibrational Entropy and Total Energy**		**MAESTRO**	**FoldX**
Lineage	**Site**	**AA Mut**	**ΔΔG (kcal mol^{-1})**	**ΔΔG (kcal mol^{-1})**	**ΔΔG$_{Vib}$ (kcal.mol^{-1} .K^{-1})**	**ΔΔG DynaMut (kcal/mol)**	**ΔΔG (kcal mol^{-1})**	**ΔΔG (kcal mol^{-1})**
B.1 B.1.1.28 P.1 P.2	21	R→I	**SUR: -0.500 (S)**	0.240 (D)	0.496 (↑M.F.)	**0.099 (S)**	**-0.256 (S)**	1.079 (D)
B.1.1.33 P.2	21	R→T	**SUR: -0.070 (S)**	-0.280 (D)	0.457 (↑M.F.)	-0.789 (D)	**-0.152 (S)**	**-0.024 (N)**
P.2	21	R→S	**SUR: 0.090 (D)**	**-0.670 (D)**	0.506 (↑M.F.)	**-0.603 (D)**	-0.065 (S)	**0.061 (N)**
B.1.1 B.1.1.28 B.1.1.29	26	P→S	**SUR: 0.000 (D)**	**-0.920 (D)**	-0.180 (↓M.F.)	**-0.179 (D)**	-0.385 (S)	**1.005 (D)**

B.1.1.33 B.1.1.44 B.1.1.7 P.1 P.1.1 P.1.2 P.2								
P.1	26	P→F	**SUR: -0.440 (S)**	-0.260 (D)	-0.310 (↓M.F.)	**0.420 (S)**	**-0.661 (S)**	0.122 (N)
P.2	26	P→L	**SUR: -0.560 (S)**	-0.220 (D)	-0.643 (↓M.F.)	**0.482 (S)**	**-0.539 (S)**	**-0.059 (N)**
P.1 P.1.2	27	A→V	**SUR: -0.620 (S)**	-0.310 (D)	-0.094 (↓M.F.)	**0.468 (S)**	**-0.276 (S)**	1.092 (D)
B.1.1.28 B.1.1.33 P.1 P.2	27	A→S	**SUR: -0.360 (S)**	-0.960 (D)	-0.219 (↓M.F.)	**0.336 (S)**	**-0.278 (S)**	0.112 (N)
B.1.1.28 P.1 P.2	29	T→I	**COR: -0.700 (S)**	**-0.200 (S)**	0.103 (↑M.F.)	-0.528 (D)	**-0.215 (S)**	2.143 (HD)
P.2	29	T→A	**COR: 0.700 (D)**	**-0.980 (D)**	0.574 (↑M.F.)	**-1.260 (D)**	-0.074 (S)	**1.936 (HD)**
B.1.1.28 B.1.1.33 B.1.195 P.1 P.2	49	H→Y	**SUR: -0.550 (S)**	-0.900 (D)	-2.809 (↓M.F.)	**1.987 (S)**	**-0.356 (S)**	**-1.303 (S)**
B.1.1.33 P.2	63	T→N	**SUR: 0.390 (D)**	-0.430 (S)	0.123 (↑M.F.)	**-0.410 (D)**	-0.169 (S)	**0.727 (SD)**
B.1.1.28 P.2	63	T→E	SUR: 0.830 (D)	**-0.370 (S)**	0.117 (↑M.F.)	**0.003 (S)**	**-0.196 (S)**	**-0.682 (SS)**
B.1.1.28 B.1.1.7 B.1.195 P.1 P.2	67	A→S	**COR: 0.170 (D)**	**-0.680 (D)**	-0.246 (↓M.F.)	**-0.927 (D)**	-0.079 (S)	**0.081 (N)**
P.1	67	A→P	**COR: -0.440 (S)**	-0.550 (D)	-0.100 (↓M.F.)	**0.046 (S)**	**-0.077 (S)**	3.173 (HD)
B.1.1 B.1.1.28 B.1.1.7 B.1.525 P.1 P.2	67	A→V	**COR: -0.130 (S)**	**0.010 (S)**	-0.727 (↓M.F.)	**1.201 (S)**	**-0.224 (S)**	**-0.071 (N)**
B.1.1.33 P.1 P.2	78	R→S	**COR: 0.180 (D)**	**-0.600 (D)**	0.935 (↑M.F.)	**-1.390 (D)**	**0.003 (D)**	**0.706 (SD)**
P.1 P.2	78	R→M	**COR: -0.390 (S)**	**0.100 (S)**	0.420 (↑M.F.)	-1.230 (D)	**-0.139 (S)**	0.565 (SD)
P.1	78	R→T	COR: -0.180 (S)	**-0.290 (D)**	0.904 (↑M.F.)	**-1.828 (D)**	-0.151 (S)	**1.178 (D)**
B.1.1.28 B.1.1.33 B.1.1.7 B.1.221	98	S→F	**SUR: -0.050 (S)**	**0.690 (S)**	-0.865 (↓M.F.)	**1.010 (S)**	**-0.464 (S)**	7.957 (HD)

P.1 P.1.2								
P.1	98	S→P	**SUR: -0.150 (S)**	**0.100 (S)**	-0.139 (↓M.F.)	**0.290 (S)**	**-0.197 (S)**	5.168 (HD)
B.1.1.28 B.1.1.33 P.1 P.1.1 P.1.2 P.2	138	D→Y	**COR: -1.270 (S)**	0.620 (D)	-0.719 (↓M.F.)	**1.699 (S)**	**-0.435 (S)**	0.799 (SD)
B.1.1.28 B.1.1.33 B.1.1.7 P.2	138	D→H	**COR: -0.560 (S)**	-1.310 (D)	-0.807 (↓M.F.)	**0.737 (S)**	**-0.318 (S)**	1.673 (D)
AV.1 B.1.617.2	142	G→D	**COR: 0.010 (D)**	**-1.130 (D)**	-0.191 (↓M.F.)	**-0.468 (D)**	-0.128 (S)	**2.848 (HD)**
B.1.234 P.2	142	G→S	COR: -0.110 (S)	**-1.260 (D)**	-0.215 (↓M.F.)	**-0.711 (D)**	-0.184 (S)	**0.069 (N)**
B.1.1.28 B.1.1.33 P.2 P.4	142	G→V	**COR: -0.900 (S)**	**-0.020 (S)**	-0.193 (↓M.F.)	**0.479 (S)**	**-0.337 (S)**	2.422 (HD)
B.1.1 B.1.1.28 B.1.1.33 B.1.177.32 B.1.177.52 B.1.617.2 P.1 P.2	222	A→V	**COR: -0.960 (S)**	**0.070 (S)**	-0.531 (↓M.F.)	**1.628 (S)**	**-0.270 (S)**	**-0,090 (N)**
P.1	222	A→S	**COR: -0.020 (S)**	-0.930 (D)	-0.252 (↓M.F.)	**0.549 (S)**	**-0.103 (S)**	0.797 (SD)
P.1	222	A→P	COR: 0.120 (D)	**-0.410 (S)**	0.016 (↑M.F.)	**0.426 (S)**	**-0.175 (S)**	**-1.427 (S)**
B.1.1 B.1.1.28 B.1.351 P.2	262	A→D	SUR: 0.630 (D)	-0.720 (D)	-0.744 (↓M.F.)	**0.100 (S)**	**-0.313 (S)**	**-0.543 (SS)**
B.1.1.28 B.1.1.33 B.1.332	262	A→G	SUR: 0.480 (D)	-1.210 (D)	-0.129 (↓M.F.)	**1.287 (S)**	**-0.133 (S)**	**-0.120 (N)**
P.1 P.1.2 P.2	262	A→S	**SUR: -0.040 (S)**	-0.400 (D)	-0.471 (↓M.F.)	**0.227 (S)**	**-0.197 (S)**	0.386 (N)
B.1.1.33 P.1 P.2	367	V→F	SUR: 0.110 (D)	-1.270 (D)	0.027 (↑M.F.)	**0.112 (S)**	**-0.564 (S)**	**-0.133 (N)**
B.1.1.28 N.9	367	V→I	**SUR: 0.100 (D)**	**-0.670 (D)**	0.109 (↑M.F.)	**-0.168 (D)**	-0.394 (S)	-0.709 (SS)
B.1.1.33	367	V→L	**SUR: 0.190 (D)**	**-0.810 (D)**	0.025 (↑M.F.)	0.040 (S)	-0.342 (S)	**0.052 (N)**
B.1.1.28 P.1	417	K→T	SUR: -0.220 (S)	**-1.220 (D)**	0.506 (↑M.F.)	**-1.022 (D)**	-0.072 (S)	**0.896 (D)**

P.1.1 P.1.2 P.2								
P.1.2	417	K→M	**SUR: -0.370 (S)**	-0.590 (D)	0.293 (↑M.F.)	-0.198 (D)	**-0.192 (S)**	**-1.410 (S)**
B.1.1.28 B.1.351	417	K→N	**SUR: 0.160 (D)**	**-1.430 (D)**	0.635 (↑M.F.)	**-0.699 (D)**	-0.068 (S)	**0.209 (N)**
P.1 P.1.2 P.2	483	V→F	**SUR: -0.000 (S)**	-0.910 (D)	0.066 (↑M.F.)	-0.393 (D)	**-0.552 (S)**	**-0.575 (SS)**
B.1.1 B.1.1.28 B.1.351 B.1.525 N.10 N.9 P.1 P.1.1 P.1.2 P.2	484	E→K	**SUR: -0.160 (S)**	-0.460 (D)	0.246 (↑M.F.)	**0.337 (S)**	**-0.433 (S)**	**-0.100 (N)**
B.1.1.33	484	E→D	**SUR: 0.190 (D)**	**-0.160 (D)**	0.227 (↑M.F.)	**-0.460 (D)**	-0.321 (S)	-0.612 (SS)
B.1 B.1.1.28 P.5	484	E→Q	**SUR: -0.090 (S)**	-0.330 (D)	0.254 (↑M.F.)	-0.285 (D)	**-0.362 (S)**	**-0.066 (N)**
AV.1 B.1.1 B.1.1.28 B.1.1.7 B.1.351 P.1 P.1.1 P.1.2 P.2	501	N→Y	**COR: -0.890 (S)**	**-0.180 (S)**	-0.135 (↓M.F.)	-0.565 (D)	**-0.365 (S)**	3.786 (HD)
B.1 B.1.1.28 B.1.1.33 P.5	501	N→T	**COR: -0.560 (S)**	**-0.300 (S)**	0.231 (↑M.F.)	-0.954 (D)	**-0.074 (S)**	0.979 (SD)
B.1.1 B.1.1.28 B.1.1.33 P.1	553	T→I	SUR: 0.210 (D)	-0.670 (D)	0.201 (↑M.F.)	**0.327 (S)**	**-0.226 (S)**	**-1.211 (S)**
B.1.1 B.1.1.7	570	A→D	**SUR: 0.450 (D)**	**-0.590 (D)**	0.037 (↑M.F.)	0.183 (S)	-0.096 (S)	**2.939 (HD)**
B.1.1.28 P.1	570	A→S	**SUR: -0.090 (S)**	-0.530 (D)	0.035 (↑M.F.)	**0.118 (S)**	**-0.139 (S)**	0.301 (N)
P.1	570	A→V	**SUR: 0.110 (D)**	**-0.340 (D)**	0.051 (↑M.F.)	0.067 (S)	-0.176 (S)	**0.116 (N)**
B.1.1.28 B.1.1.7 P.1 P.1.2 P.2	572	T→I	SUR: 0.080 (D)	-0.790 (D)	-0.293 (↓M.F.)	**0.764 (S)**	**-0.307 (S)**	**-0.922 (S)**
B.1.1.33 P.2	572	T→N	**SUR: -0.200 (S)**	-0.690 (D)	0.008 (↑M.F.)	-0.336 (D)	**-0.235 (S)**	**-0.542 (SS)**

B.1.1.28 P.1 P.2	585	L→F	COR: -1.380 (S)	**-1.420 (D)**	-0.443 (↓M.F.)	**-0.574 (D)**	-0.121 (S)	**1.563 (D)**
P.1	585	L→V	**COR: 0.880 (D)**	**-1.430 (D)**	0.355 (↑M.F.)	**-0.619 (D)**	-0.065 (S)	**1.512 (D)**
B.1 and derived	614	D→G	**SUR: -0.320 (S)**	**-0.230 (S)**	0.557 (↑M.F.)	-0.082 (D)	0.013 (D)	**-0.238 (SS)**
P.1	614	D→S	**SUR: -0.260 (S)**	**0.110 (S)**	0.244 (↑M.F.)	-0.981 (D)	**-0.137 (S)**	0.832 (SD)
B.1.1.33 N.5 P.1 P.4	653	A→V	**COR: -0.630 (S)**	-0.460 (D)	-0.549 (↓M.F.)	**0.615 (S)**	**-0.269 (S)**	0.247 (N)
P.1	653	A→S	**COR: 0.380 (D)**	**-1.240 (D)**	0.082 (↑M.F.)	**-1.188 (D)**	-0.014 (S)	**1.631 (D)**
B.1.1.28 B.1.1.33 B.1.1.464 P.1 P.2	689	S→I	**SUR: -0.420 (S)**	**-0.140 (S)**	-4.514 (↓M.F.)	**1.394 (S)**	**-0.607 (S)**	**-0.105 (N)**
B.1.1.33	689	S→G	**SUR: -0.070 (S)**	-0.270 (D)	-4.426 (↓M.F.)	**1.526 (S)**	**-0.421 (S)**	**-0.049 (N)**
B.1.1.28	701	A→S	**SUR: 0.050 (D)**	**-0.320 (D)**	0.002 (↑M.F.)	**-0.164 (D)**	-0.387 (S)	**0.382 (N)**
B.1.1.33 B.1.1.7 B.1.351 P.1 P.2	701	A→V	**SUR: -0.420 (S)**	**-0.040 (S)**	0.066 (↑M.F.)	**0.034 (S)**	**-0.360 (S)**	0.484 (SD)
B.1.1.28 P.1 P.2	747	T→I	**SUR: -0.130 (S)**	-0.290 (D)	0.002 (↑M.F.)	**0.146 (S)**	**-0.528 (S)**	0.042 (N)
P.1 P.2	747	T→A	**SUR: 0.070 (D)**	**-0.640 (D)**	0.018 (↑M.F.)	0.159 (S)	-0.337 (S)	**0.739 (SD)**
B.1.1.28 B.1.1.7 N.9 P.2	769	G→V	SUR: 0.330 (D)	0.070 (D)	-0.072 (↓M.F.)	**0.103 (S)**	**-0.510 (S)**	**-0.777 (SS)**
B.1.1.378 B.1.195	769	G→R	SUR: 0.460 (D)	-0.130 (D)	-0.115 (↓M.F.)	**0.672 (S)**	**-0.809 (S)**	**-1.045 (S)**
B.1.1.28 B.1.1.7 P.1 P.2	845	A→S	**SUR: -0.340 (S)**	-0.730 (D)	-0.002 (↓M.F.)	**0.031 (S)**	**-0.117 (S)**	0.360 (N)
P.1	845	A→V	**SUR: 0.250 (D)**	**-0.260 (D)**	-0.127 (↓M.F.)	0.389 (S)	**-0.228 (S)**	**1.752 (D)**
B.1.1.28 P.1 P.2	846	A→S	**SUR: 0.420 (D)**	**-0.450 (D)**	-0.871 (↓M.F.)	**-0.198 (D)**	-0.223 (S)	-0.528 (SS)
B.1.1 B.1.1.7 P.2	846	A→V	**SUR: -0.730 (S)**	-0.220 (D)	-0.332 (↓M.F.)	**0.230 (S)**	**-0.308 (S)**	0.683 (SD)

B.1.1.28 B.1.1.33 P.1 P.1.1 P.1.2 P.2	1027	T→I	SUR: 0.460 (D)	-0.310 (D)	0.221 (↑M.F.)	**0.242 (S)**	**-0.359 (S)**	**-0.828 (SS)**
B.1.1.28 N.9 P.1 P.2	1078	A→S	COR: 0.090 (D)	-1.020 (D)	-0.043 (↓M.F.)	**0.004 (S)**	**-0.035 (S)**	**-0.621 (SS)**
N.9	1078	A→V	**COR: -0.520 (S)**	-0.530 (D)	-0.273 (↓M.F.)	**1.684 (S)**	**-0.223 (S)**	**-0.209 (N)**
B.1.1.33	1078	A→T	COR: 0.210 (D)	-1.010 (D)	-0.297 (↓M.F.)	**0.233 (S)**	**-0.167 (S)**	**-0.344 (N)**
B.1.1 B.1.1.28 P.1	1084	D→Y	SUR: 0.020 (D)	**-0.270 (S)**	-0.043 (↓M.F.)	-0.063 (D)	**-0.616 (S)**	**-0.202 (N)**
P.1	1084	D→H	**SUR: -0.020 (S)**	**-0.410 (S)**	-0.004 (↓M.F.)	-0.171 (D)	**-0.526 (S)**	0.107 (N)
B.1.1.33 P.1	1124	G→V	**SUR: -0.370 (S)**	-0.660 (D)	-0.234 (↓M.F.)	**0.953 (S)**	**-0.411 (S)**	5.895 (HD)
B.1.1.7	1124	G→A	**SUR: -0.330 (S)**	-0.790 (D)	-0.129 (↓M.F.)	**0.729 (S)**	**-0.201 (S)**	3.101 (HD)
B.1.1.33 P.1	1133	V→F	**SUR: 0.370 (D)**	**-1.110 (D)**	-0.205 (↓M.F.)	**-0.004 (D)**	-0.329 (S)	-0.731 (SS)

Sites with no covered position by the structural analysis: 5, 12, 75, 76, 681, 684, 688, 1167, 1176, 1219, 1228, 1260, and 1264. Effect: (SS) Slightly Stabilizing; (S) Stabilizing; (N) Neutral; (SD) Slightly Destabilizing; (D) Destabilizing; and (HD) Highly Destabilizing. Molecular Flexibility: (↓M.F.) Decreased Molecular Flexibility; (↑M.F.) Increased Molecular Flexibility. See methods (Spike structural stability) for the thresholds of each category. For FoldX, positive values are considered destabilizing, while negative values are considered stabilizing. The majority consensus among methods are highlighted in bold.

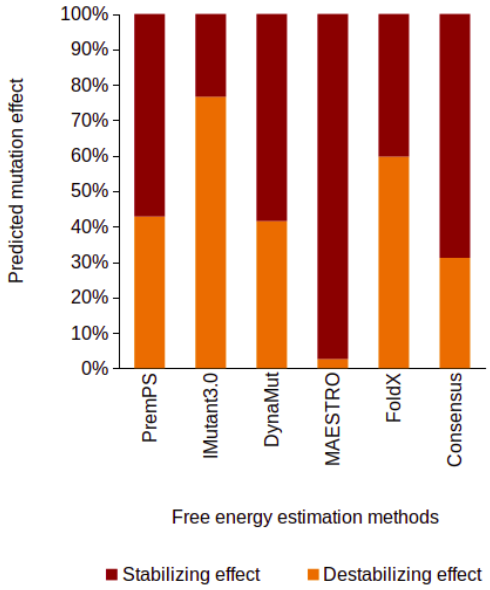

Figure 4. Predicted mutation effect on the positively selected sites according to the tested methods (PremPS, iMutant3.0, DynaMut, MAESTRO, and FoldX) and the majority consensus results (energy trend estimated by three or more methods).

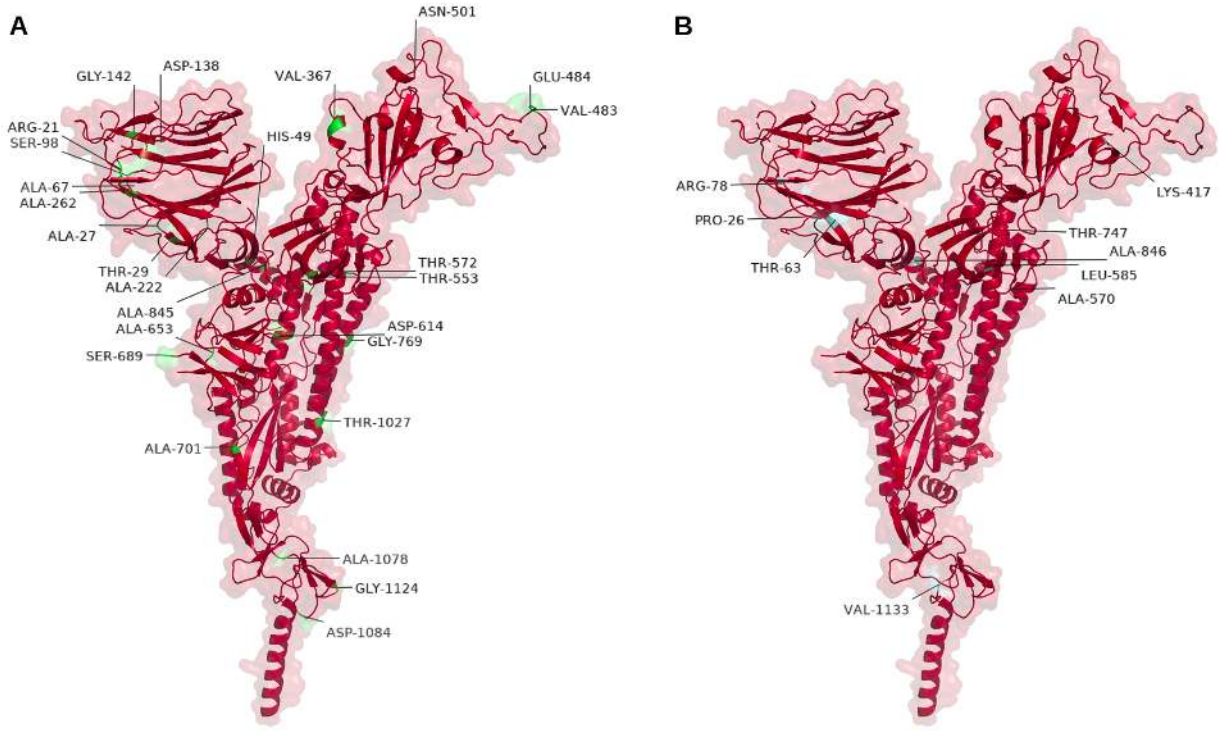

Figure 5. Identification of the stabilizing and destabilizing mutations in the prefusion spike protein structure (6XR8, chain A). (A) Stabilizing mutations (highlighted in green and labeled)

related to the positively selected sites, according to the consensus majority analysis. (B) Destabilizing mutations (highlighted in cyan and labeled) related to the positively selected sites, according to the majority consensus analysis. For visualization purposes, sites with multiple possible mutations were represented by the prevalent one and its respective effect (stabilizing/destabilizing).

Spike RBD-ACE2 structural modelling

In total, 30 theoretical models with 194 amino acid length for the spike RBD - ACE2 protein complexes were obtained using PDB ID 6M0J as template, being five models generated for each lineage (reference, P.1, P.2, C.37, B.1.1.7, and P.2+452). The model quality was evaluated by analysis of DOPE score and through PROCHECK/VERIFY3D parameters (Supplementary File 5). The models covered the region between the amino acid positions 333 and 526 of the reference spike protein sequence (relative to the genome NC_045512.2) and from five SARS-CoV-2 lineages: C.37 (L452Q + F490S), B.1.1.7 (N501Y), P.1 (K417T + E484K + N501Y), P.2 (E484K), and P.2+452 (E484K + L452V) (Figure 6). The final selected models for the reference as well as for the lineages C.37, B.1.1.7, P.1, P.2, and P.2+452 had 88.7 up to 89.9% of the residues in favoured regions and 10.1 up to 11.3% of the residues in allowed regions, which indicated the good model quality. There were no residues in generously allowed or disallowed regions (Figure S1). The overall G-Factor values ranged between -0.06 and -0.12. For all selected models, the compatibility of the atomic model with the amino acid sequence was defined with 100% of the residues achieving an average 3D-1D score ≥ 0.2 in the VERIFY3D analysis.

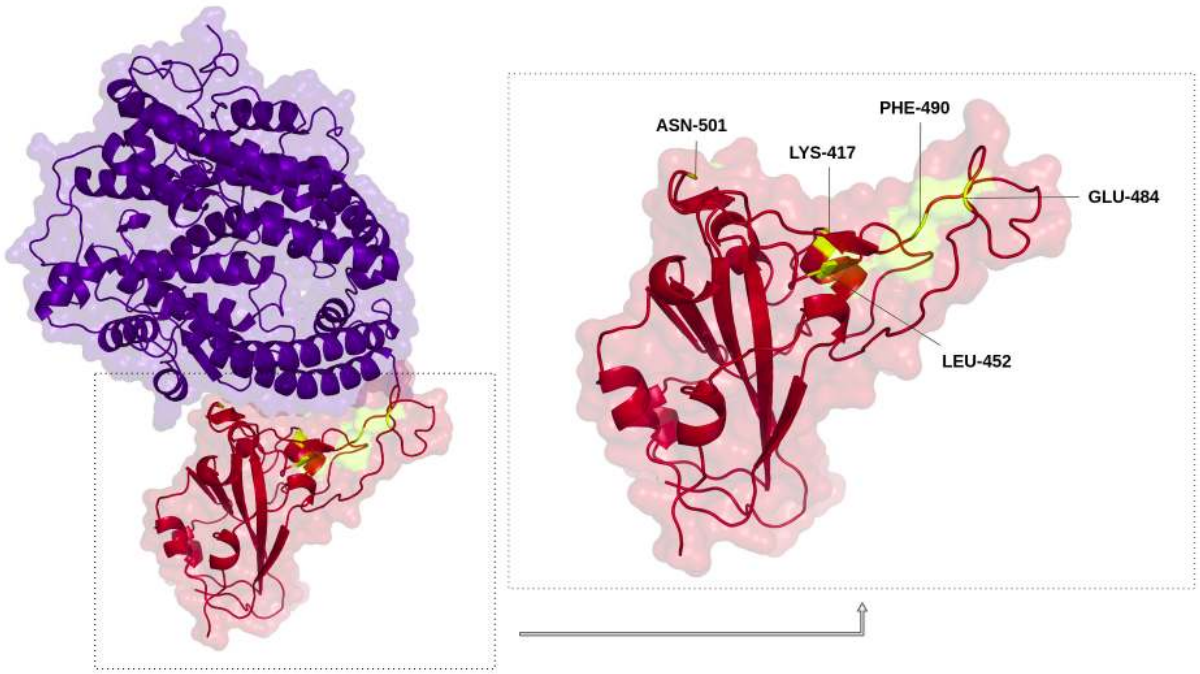

Figure 6. Spike Receptor Binding Domain (RBD) from NC_045512.2 reference genome complexed with human Angiotensin Converting Enzyme 2 (hACE2). The ACE2 structure is colored in blue. The localization of the mutated residues for the modelled RBD structures from lineages C.37 (L452, F490), B.1.1.7 (N501), P.1 (K417, E484, N501), P.2 (E484), and P.2+452 (L452, E484) are highlighted in yellow and labeled.

Molecular dynamics and binding free energy estimation

The analysis of the RMSD, center of mass and minimum distances indicated that all systems were structurally stable, in the considered time window, with overall conservation of the arrangement of the complex, as can also be seen by visual inspection of the snapshots (Figures 7 and 8). The lowest changes were found in the reference and lineages P.1 and C.37, meaning higher structural stability. Some structural changes in spike were found in the P.2 and B.1.1.7 lineages, whereas for P.2+452 and B.1.1.7 the structural changes were also noticeable in relation to the ACE2 molecule. For all systems, the spike-hACE2 interaction was characterized by the presence of many close contacts, which is reflected in a substantial interface area, and many hydrogen bonds (Figure 9 and Table 4).

Despite the distance plots with similar results, indicating that the relative position between the spike and ACE2 is not altered during the simulation time, the lineages P.2+452, C.37 and B.1.1.7 showed a small alteration in the distance, with a slight approximation between spike and ACE2 (Figure S2). In the reference system, there were on average 9.51 ± 2.09 hydrogen bonds and an interface area 10.15 ± 0.55 nm^2. The systems P.2, P.2+452, and C.37 displayed a higher number of hydrogen bonds and interface area than the reference, whereas the lineage P.1 showed larger interface area but less hydrogen bonds and the lineage B.1.1.7 showed smaller interface area and number of hydrogen bonds. About the number and temporal evolution of the hydrogen bonds, the order followed: P.2 > P.2+452 > C.37 > reference > P.1 > B.1.1.7. The intensity of the hydrogen bonds in the lineages P.2 and P.2+452 is reflected in the interaction intensity with ACE2. As for the contact area between spike and ACE2, proportional to the intensity of the van der Waals interactions (also reflected in the interaction intensity), the order followed with a slight change: P.2+452 > P.2 > P.1 > C.37 > reference > B.1.1.7 (Figure 9 and Table 5).

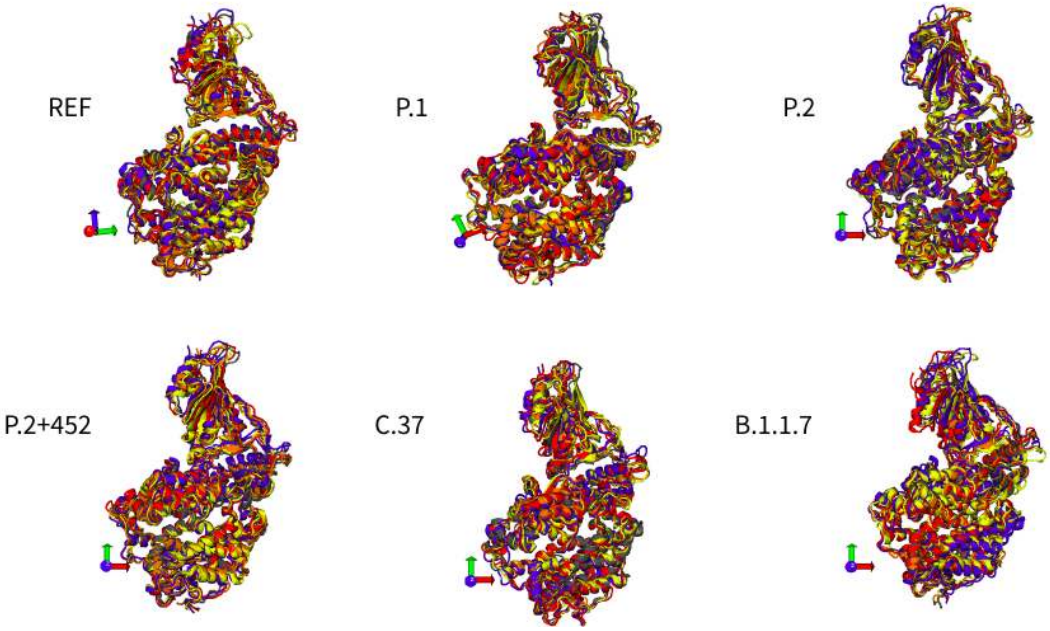

Figure 7. Snapshots from molecular dynamics (MD) simulations in 0, 50, 100, 150, and 200 ns for the spike RBD - ACE2 protein complexes related to the reference (REF) and lineages P.1, P.2, P.2+452, C.37, and B.1.1.7.

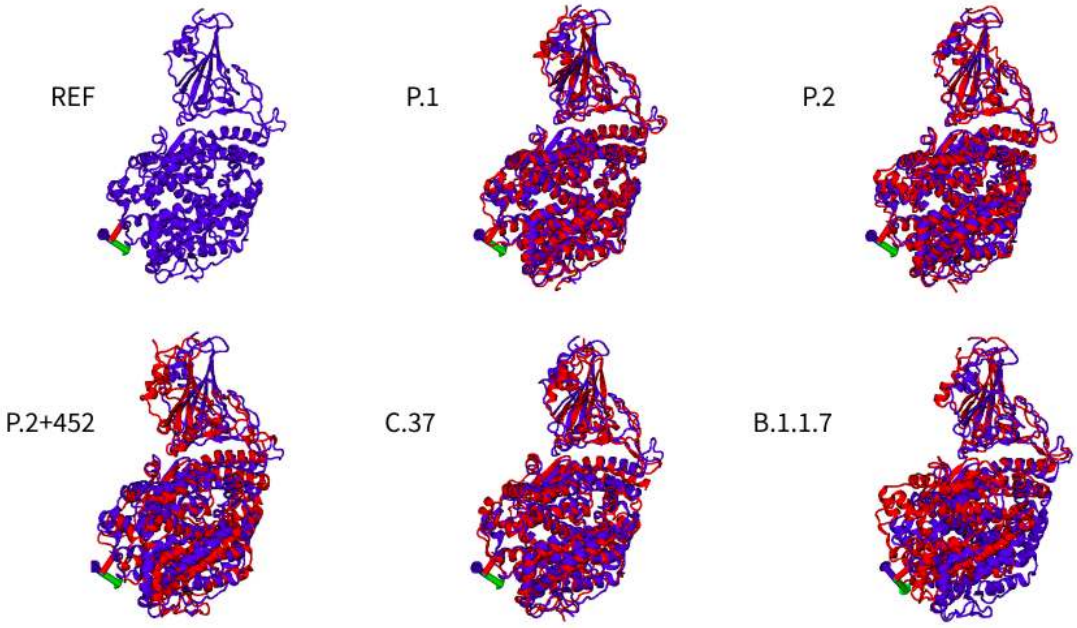

Figure 8. Snapshots from molecular dynamics (MD) simulations - the final configuration fitted in relation to the reference (REF) - for the spike RBD - ACE2 protein complexes belonging to lineages P.1, P.2, P.2+452, C.37, and B.1.1.7.

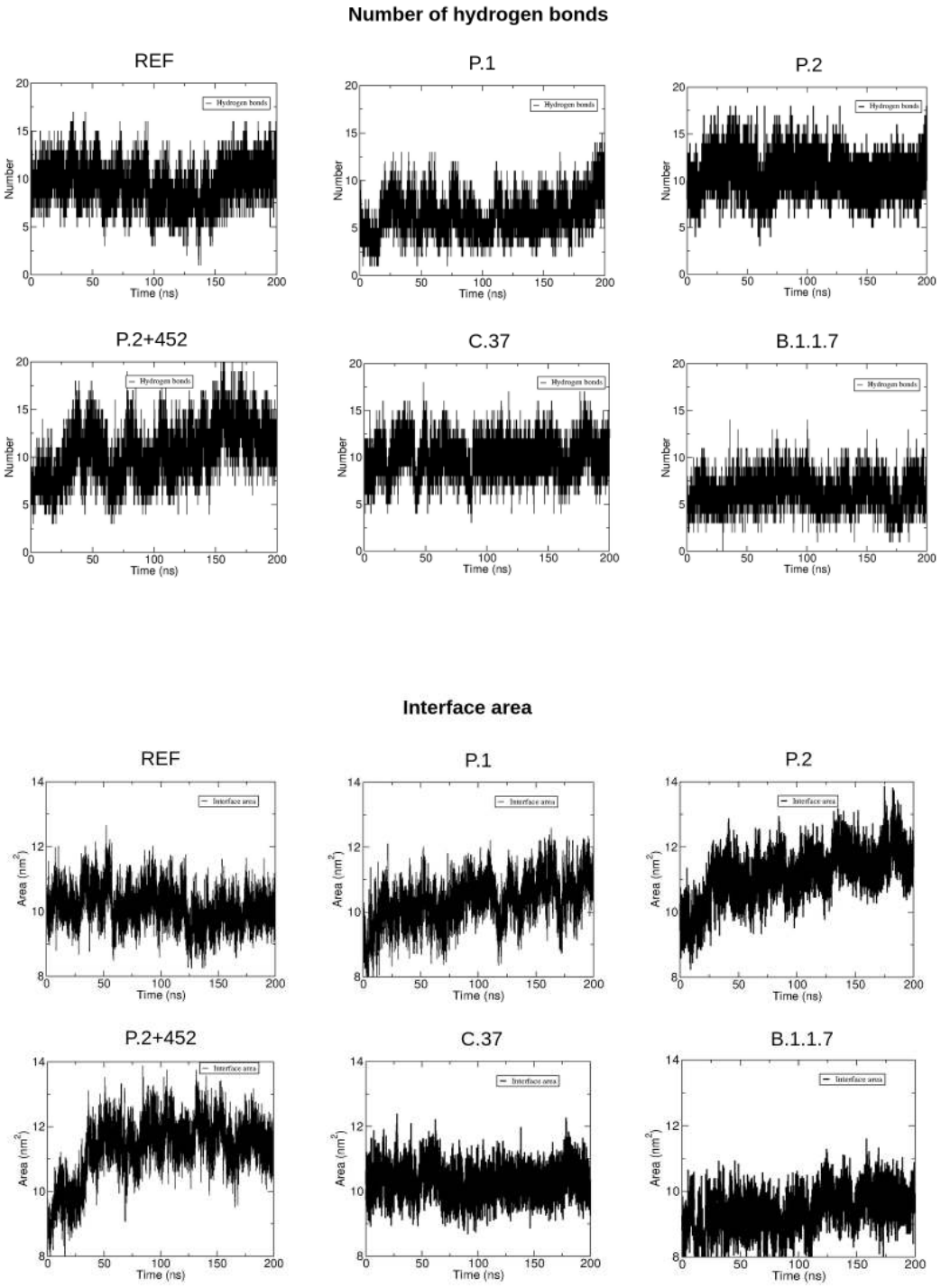

Figure 9. Number of hydrogen bonds and interface area for the spike RBD - ACE2 protein complexes related to the reference (REF) and lineages P.1, P.2, P.2+452, C.37, and B.1.1.7.

Table 4. Interface area of the complexes (the half of the sum of solvent-accessible areas of spike and ACE2 minus solvent-accessible area of the complex) and number of hydrogen bonds, calculated as average over the MD simulations.

Complex	Interface area (nm^2)	Average number of hydrogen bonds
REF	10.15 ± 0.55	9.51 ± 2.09
P.1	10.33 ± 0.69	6.33 ± 2.01
P.2	11.13 ± 0.77	10.64 ± 1.91
P.2+452	11.31 ± 0.89	10.47 ± 2.70
C.37	10.31 ± 0.46	9.98 ± 1.85
B.1.1.7	9.40 ± 0.58	6.26 ± 1.66

In all cases, the binding free energy of the mutant spikes was found to be more negative (stronger binding) than the native one and the order is roughly similar to the intensity of interface area or hydrogen bonds: P.2+452 > P.2 > P.1 > B.1.1.7 > C.37 > reference. Specifically, for P.2+452 and P.2, the binding intensity was significantly stronger than the reference, while for B.1.1.7, C.37 and reference the interaction intensity was almost the same (Table 5).

In relation to the hotspots for the RBD-ACE2 interaction, the residue distances for the wild-type (reference) and the mutated sites were calculated (Supplementary Figures S3-S7). The spike K417 interacts with hACE2 residues T27, D30, K31, and H34 in the wild type, while the mutation for T417 in P.1 maintains the same contacts. In the P.2 lineage, K417 is affected by the E484K mutation, changing the residue interactions to the hACE2 D30, N33, H34, and P389, which is the same profile identified in B.1.1.7, despite the higher distance variation during the simulation time. The impact of the E484K + L452V combination in P.2+452 changes the residue interaction of K417 to D30, H34, Q388, and P389. For the C.37 lineage, the effect of L452Q and F490S modified the residue interaction of K417 to T27, D30, H34, and P389. The residue N501 in the wild type RBD-ACE2 complex interacts with L351, G352, K353, G354, D355, and Y41. This same pattern is kept by P.1, with a slight reduction in the Y501-K353 distance. N501 of P.2, P.2+452 and C.37 lineages presented the same ACE2 interactions. However, C.37 showed a

slight increase in the distance value for F356 and D355 in the middle and end of the simulation time. Finally, Y501 in B.1.1.7 showed the highest variation, creating a new interaction with hACE2 residue H34.

Table 5. Binding free energies and their components, calculated using MM/PBSA, for the interaction between ACE2 and the lineages a) REF, b) P.1, c) P.2, d) P.2+452, e) C.37, and f) B.1.1.7.

Complex	Van der Waals Energy (kJ/mol)	Electrostatic Energy (kJ/mol)	Polar Solvation Energy (kJ/mol)	SASA Energy (kJ/mol)	Binding Energy (kJ/mol)
REF	-406.772 +/- 2.081	-1405.564 +/- 7.144	724.141 +/- 10.657	-45.611 +/- 0.286	**-1134.283 +/- 10.350**
P.1	-417.644 +/- 2.970	-1627.348 +/- 7.132	533.145 +/- 11.232	-45.812 +/- 0.302	**-1558.114 +/- 10.124**
P.2	-448.576 +/- 3.682	-2348.699 +/- 7.801	811.214 +/- 11.373	-49.739 +/- 0.325	**-2035.611 +/- 10.218**
P.2+452	-437.065 +/- 3.066	-2494.514 +/- 8.456	911.602 +/- 11.865	-50.055 +/- 0.351	**-2069.977 +/- 9.690**
C.37	-417.849 +/- 2.285	-1412.224 +/- 8.546	740.556 +/- 12.203	-46.253 +/- 0.257	**-1135.494 +/- 9.501**
B.1.1.7	-374.295 +/- 2.474	-1295.292 +/- 5.377	571.656 +/- 10.855	-41.636 +/- 0.274	**-1140.138 +/- 11.183**

In relation to the magnitude of interaction, all three top-ranked lineages bear the same E484K mutation, which yielded a particularly strong binding, remarkably stronger than in the native case. Indeed, the analysis of the contribution of the residues to the free energy of binding showed that the mutation of the negatively charged residue GLU 484 to the positive charged residue LYS 484 yielded a substantial negative free energy contribution to the stabilization of the complex. The mutation in the residue 501 (ASN to TYR, both polar non charged residues) presented in the lineages P.1 and B.1.1.7 can also contribute to the stabilization of the complex, but much less than E484K (Table 6).

Table 6. Contribution of mutations to the Binding Free Energy changes (kJ/mol).

REF		P.1		P.2		P.2+452		C.37		B.1.1.7	
Lys417	-236.4	**Thr417**	**-4.6**	Lys417	-239,9	Lys417	-236,8	Lys417	-241,3	Lys417	-239,8
Leu452	-1.6	Leu452	-2.0	Leu452	-1.6	**Val452**	**0**	**Gln452**	**-3.1**	Leu452	-1.8
Glu484	249.5	**Lys484**	**-211.4**	**Lys484**	**-218.0**	**Lys484**	-234.5	Glu484	228.4	Glu484	224.3
Phe490	0.4	Phe490	1.2	Phe490	3.4	Phe490	0.4	**Ser490**	**3.2**	Phe490	1.3
Asn501	-12.0	**Tyr501**	**-15.7**	Asn501	-12.5	Asn501	-10.7	Asn501	-12.6	**Tyr501**	-15.5

DISCUSSION

The spike protein is one of the most rapidly evolving regions in the SARS-CoV-2 genome, with slightly different evolution rates according to the clades, emerging in the beginning and middle 2020 (Pereson et al., 2020). The emergence of the P.1 lineage, at the end 2020, with a possibly increased rate of the molecular evolution, represents an important change in the SARS-CoV-2 evolutionary history (Faria et al., 2021). That is the main reason why lineage assignments should be carefully evaluated, since a considerable number of sequences generally described as P.2 or B.1.1.28 (e.g.) presents a mutation set consistent with the P.1 lineage and are grouped in the same monophyletic clade as well. A systematic and detailed phylogenomic analysis should be conducted in order to better understand and classify new SARS-CoV-2 genomes. The prevalence of P.1 containing mutations in the set of sites indicated to be under adaptive selective pressure may suggest the improvement of viral fitness in this lineage. In fact, a high number of mutations were found in P.1 variants, besides the 10 lineage-defining mutations (https://outbreak.info/situation-reports?pango=P.1). Interestingly, eight of them were already found to be positively selected in the study of Faria and colleagues (2021), using a small set of P.1 and P.2 sequences. However, our analyses included 2,901 unique spike sequences from several lineages and identified pervasive adaptive selection signatures in sites 26, 138, 417, 484, 501, and 1,027, whose mutations are found occurring in more than nine Brazilian lineages. Therefore, the possible change of a neutral genetic drift to a

strong selective pressure may be related to the P.1 mutational advantage facing the population-level immunity, according to Van Egeren et al. (2021). In this regard, it is interesting to note that P.1 has emerged in a region with up to 75% demonstrated previous seroprevalence of SARS-CoV-2. Similarly, lineage-defining mutated sites from B.1.1.7 (570 and 681) and C.37 lineages (75 and 76) were also identified in the positive selection tests, strongly suggesting that similar phenomena of selective pressure have been the crucial in the emergence of others lineages as well.

According to Shindyalov and colleagues (1994), mutations in certain positions are fixed in a correlated manner along the evolution. This may be the result of structural or functional constraints imposed for the maintenance of the protein integrity and stability. Although close chain neighbours have higher probabilities to mutate in a conditional way, even distant structural positions can follow this pattern through a number of different mechanisms. Coevolution may be the result of compensatory, fitness recovery interactions. Alternatively, mutations that enhance infectivity, such as N501Y, may increase the chance of further viral evolution. Recently, for example, the acquisition of type I interferon resistance, a seemingly key factor for pathogenesis in SARS-CoV, has been found in the newer S harboring mutant lineages (Guo et al., 2021; Kim and Shin, 2021). A third potential cause of viral coevolution through mutation linkage could be viral immune escape. The H69-V70 deletion, for instance, present in NTD from the B.1.1.7 lineage, abolishes an important immunogenic loop by allowing the inward movement of the amino acids 71-75 from that domain. In this work, the analysis of coevolution with the Bayesian Graphical Model showed that 28 pairs of sites (including the couple 484 - 501) can be statistically related with a posterior probability ≥ 0.8, indicating that these sites are probably not conditionally independent. The co-occurrence of E484K and N501Y is an interesting fact, since N501Y emerged independently in P.1 and B.1.351 lineages (Winger and Caspari, 2021), as well E484K (Ferrareze et al., 2021).

We have demonstrated that, individually, some mutations may stabilize or destabilize the protein structure, since their occurrence triggers different effects on energy balances and potentially affects viral fitness. As shown by the result of our majority consensus analysis and demonstrated by Pucci and Rooman (2021), mutations such as A222V (B.1.617.2), N501Y (B.1.1.7, P.1, B.1.351), and D614G generate a stabilizing effect in the spike structure. The D614G, specifically, reduces furin cleavage, lowering the risk of premature S1 shedding (Winger and Caspari, 2021). About the set of P.1 lineage-defining mutations found in the RBD region, the prevalent E484K substitution, located in the loop region outside the direct hACE2 interface, also stabilizes the spike protein structure, reducing the unfolding free energy changes. This effect was also predicted for E484Q, identified in Brazilian SARS-CoV-2 viruses from B.1, B.1.1.28, and P.5 lineages. However, E484D (B.1.1.33), represented by a single sequence in the genome set (and 44 sequences worldwide https://outbreak.info/situation-reports?pango&muts=S%3AE484D) seems to destabilize the spike structure. The analysis of the multiple substitutions potentially fixed by sites under adaptive selection with the strongest evidence (i.e., those 17 sites identified by FUBAR, FEL, and SLAC) showed that sites such as 138 (P.1, P.1.1, P.1.2, P.2), 222 (B.1.617.2, P.1, P.2), 262 (P.1, P.1.2, P.2, B.1.351), 572 (B.1.1.7, P.1, P.1.2, P.2), 614, and 845 (B.1.1.7, P.1, P.2) presented a concordant stabilizing trend for all the evaluated substitutions. However, the high frequency observed for the mutations D138Y (55.04%) and D614G (99.14%) can not be extended to any other site substitutions. In fact, considering 17 sites under positive selection identified by the three HyPhy methods, only six had substitution frequencies above 1% in the whole genome set (138, 484, 614, 681, 845, and 1,176).

The association of N501Y with D614G (B.1.1.7, P.1, B.1.351) improves the thermostability as well as increases the binding affinity with hACE2 (Winger and Caspari, 2021). Several studies (Chen et al., 2021; Hark Gan et al., 2021) have demonstrated the increased binding affinity upon the amino acid changes E484K+N501Y+K417T (P.1), superposing the

K417T effect that decreases the binding affinity. The threonine substitution at site 417 avoids the salt bridge formation with D30 in hACE2. Consequently, there is expected diminution in the binding affinity for the mutated structures (Winger and Caspari, 2021). As previously elucidated, the binding free energy between RBD and ACE2 reflects the infectivity (Qu et al., 2005). Previous analysis of the substitutions L452R, F490S, and N501Y presented relatively high binding free energy changes suggesting that these sites may generate more infectious lineages (Wang et al., 2021). Our findings corroborates the improved binding pattern found in P.1 models, showing the greater contribution of E484K mutation to the binding free energy reduction and the RBD-hACE complex stabilization in P.1 and P.2 lineages, especially in the presence of a mutated L452V site. Despite L452Q (C.37) suggesting similar properties to L452R (B.1.617.2), increasing viral infectivity (Acevedo et al., 2021), nothing is known about the valine substitution and its effect on P.2 viruses. However, a significant interaction effect with E484K was observed in this study, also with indication of structural changes in the ACE2 molecule (as for B.1.1.7 model). Interestingly, the escape from neutralizing antibodies may be related to mutants with increased ACE2 binding affinities (Van Egeren et al., 2021). As shown by our molecular dynamic simulations, the comparison of the combined binding energies for the evaluated lineages indicates a major reduction in the binding free energy of those containing E484K, with only a small difference between the reference and lineages such as C.37 and B.1.1.7.

In our data, mutations such as P26S (B.1.1.7, P.1, P.1.1, P.1.2, P.2), T63N (P.2), R78S/T (P.1, P.2), K417T/N (P.1, P.1.1, P.1.2, P.2, B.1.351), A570D/V (B.1.1.7, P.1), L585F/V (P.1, P.2), T747A (P.1, P.2), A846S (P.1, P.2), and V1133F (P.1), among others, in positively selected sites, seem to impact the spike prefusion structure with a destabilizing effect. Except for P26S and K417T, which are found in 58.45% and 49.11% of the spike sequences from the genome set (n=11,078), respectively, the remaining mutations are observed in very low frequencies.

The selection of theoretically “destabilizing” or "stabilizing" mutations still remains to be completely elucidated. One possible explanation for fixation of such mutations would be the

occurrence of “copy choice” recombinations capable of selecting very quickly for a set of mutations that have coevolved for the aforementioned reasons. The other possibility would be rapid evolution mediated by lineages enhanced with greater replicative fitness, as a consequence of key RBD mutations and/or other allosteric interactions (*e.g.*, D614G). The higher viral turnover would account for a tremendous selective pressure and fast viral selection, provided the viruses find an appropriate environment for transmissions. In this regard, it is interesting to note that both B.1.351 and P.1 appeared to have emerged in a very intense selective pressure scenario. In some Amazonas State cities in Brazil, where the P.1 lineage has emerged, up to 75% exposure were found for previous first-wave lineage. Similarly, B.1.351 has emerged in South Africa in a scenario with up to 30% of the population being previously exposed to other lineages. Taken together this evidence supports the role of consecutive accumulation of mutations rather than recombinations in order to explain the emergence of multiple lineages with specific sets of mutations, characteristic of the VOC viruses. Finally, it is possible that an immunosuppressed host would be critical for selecting these viruses, since the first occurrence of important, but intrinsically “destabilizing” substitution, would prevent further evolution if the virus encounters a robust immune response. In other hand, an immunocompromised patient would allow the progressively mutated virus to evolve, eventually acquiring fitness recovery and/or escape mutants permitting transmission to normal hosts. Although this hypothesis is merely speculative at this time, previous works involving multiple genotyping of a single chronically immunosuppressed patient (Choi et al., 2020; Kemp et al., 2021) suggested that this could be a way of rapidly selecting virus harboring mutations that, on an individual basis, are “destabilizing”.

Availability of data and materials

Full tables acknowledging the authors and corresponding labs submitting sequencing data used in this study can be found in Supplementary File S1. Additional information used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare no competing interests.

Funding

Scholarships and Fellowships were supplied by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA). The funders had no role in the study design, data generation and analysis, decision to publish or the preparation of the manuscript.

Author's contributions

Patrícia A. G. Ferrareze: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Visualization, Writing - Original Draft, Writing - Review & Editing. **Ricardo A. Zimerman:** Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing. **Vinicius B. Franceschi:** Investigation, Visualization, Writing - Original Draft, Writing - Review & Editing. **Gabriel D. Caldana:** Writing - Original Draft, Writing - Review & Editing. **Paulo A. Netz:** Methodology, Software, Formal analysis, Investigation, Resources, Visualization, Writing - Original Draft, Writing - Review & Editing. **Claudia E. Thompson:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration. All authors have read and approved the manuscript.

Acknowledgements

We thank Amanda M. Mayer for suggestions in the introduction section. We also thank the administrators of the GISAID database and research groups across the world for supporting the rapid and transparent sharing of genomic data during the COVID-19 pandemic.

REFERENCES

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

Acevedo, M.L., Alonso-Palomares, L., Bustamante, A., Gaggero, A., Paredes, F., Cortés, C.P., Valiente-Echeverría, F., Soto-Rifo, R., 2021. Infectivity and immune escape of the new SARS-CoV-2 variant of interest Lambda. Cold Spring Harbor Laboratory.

Ali, F., Kasry, A., Amin, M., 2021. The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutant. Med. Drug Discov. 10, 100086. https://doi.org/10.1016/j.medidd.2021.100086

Baker, N.A., Sept, D., Joseph, S., Holst, M.J., McCammon, J.A., 2001. Electrostatics of nanosystems: Application to microtubules and the ribosome. Proc. Natl. Acad. Sci. 98, 10037–10041.

Baum, A., Fulton, B.O., Wloga, E., Copin, R., Pascal, K.E., Russo, V., Giordano, S., Lanza, K., Negron, N., Ni, M., Wei, Y., Atwal, G.S., Murphy, A.J., Stahl, N., Yancopoulos, G.D., Kyratsous, C.A., 2020. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science 369, 1014–1018. https://doi.org/10.1126/science.abd0831

Boni, M.F., Posada, D., Feldman, M.W., 2007. An Exact Nonparametric Method for Inferring Mosaic Structure in Sequence Triplets. Genetics 176, 1035–1047. https://doi.org/10.1534/genetics.106.068874

Cai, Y., Zhang, J., Xiao, T., Peng, H., Sterling, S.M., Walsh, R.M., Rawson, S., Rits-Volloch, S., Chen, B., 2020. Distinct conformational states of SARS-CoV-2 spike protein. Science 369, 1586–1592. https://doi.org/10.1126/science.abd4251

Capriotti, E., Fariselli, P., Casadio, R., 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 33, W306-310. https://doi.org/10.1093/nar/gki375

Chen, C., Boorla, V.S., Banerjee, D., Chowdhury, R., Cavener, V.S., Nissly, R.H., Gontu, A., Boyle, N.R., Vandergrift, K., Nair, M.S., Kuchipudi, S.V., Maranas, C.D., 2021. Computational prediction of the effect of amino acid changes on the binding affinity between SARS-CoV-2 spike protein and the human ACE2 receptor. bioRxiv 2021.03.24.436885. https://https://doi.org/10.1101/2021.03.24.436885

Chen, Y., Lu, H., Zhang, N., Zhu, Z., Wang, S., Li, M., 2020. PremPS: Predicting the impact of missense mutations on protein stability. PLOS Comput. Biol. 16, e1008543. https://doi.org/10.1371/journal.pcbi.1008543

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., Brigl, M., Kwon, D.S., Kanjilal, S., Richardson, E.T., Jonsson, A.H., Alter, G., Barczak, A.K., Hanage,

W.P., Yu, X.G., Gaiha, G.D., Seaman, M.S., Cernadas, M., Li, J.Z., 2020. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. N. Engl. J. Med. 383, 2291–2293. https://doi.org/10.1056/NEJMc2031364

Darden, T., York, D., Pedersen, L., 1993. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. J. Chem. Phys. 98, 10089–10092. https://doi.org/10.1063/1.464397

Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., Flouri, T., 2020. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. Mol. Biol. Evol. 37, 291–294. https://doi.org/10.1093/molbev/msz189

Davies, N.G., Jarvis, C.I., Edmunds, W.J., Jewell, N.P., Diaz-Ordaz, K., Keogh, R.H., 2021. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. Nature 593, 270–274. https://doi.org/10.1038/s41586-021-03426-1

Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., Kollman, P., 2003. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J. Comput. Chem. 24, 1999–2012. https://doi.org/10.1002/jcc.10349

Eisenberg, D., Lüthy, R., Bowie, J.U., 1997. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 277, 396–404. https://doi.org/10.1016/s0076-6879(97)77022-8

Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., Scharf, M., 1995. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J. Comput. Chem. 16, 273–284. https://doi.org/10.1002/jcc.540160303

Faria, N., Mellan, T.A., Whittaker, C., Claro, I.M., Candido, D. da S., Mishra, S., Crispim, M.A.E., Sales, F.C.S., Hawryluk, I., McCrone, J.T., Hulswit, R.J.G., Franco, L.A.M., Ramundo, M.S., Jesus, J.G. de, Andrade, P.S., Coletti, T.M., Ferreira, G.M., Silva, C.A.M., Manuli, E.R., Pereira, R.H.M., Peixoto, P.S., Kraemer, M.U.G., Gaburo, N., Camilo, C. da C., Hoeltgebaum, H., Souza, W.M., Rocha, E.C., Souza, L.M. de, Pinho, M.C. de, Araujo, L.J.T., Malta, F.S.V., Lima, A.B. de, Silva, J. do P., Zauli, D.A.G., Ferreira, A.C. de S., Schnekenberg, R.P., Laydon, D.J., Walker, P.G.T., Schlüter, H.M., Santos, A.L.P. dos, Vidal, M.S., Caro, V.S.D., Filho, R.M.F., Santos, H.M. dos, Aguiar, R.S., Proença-Modena, J.L., Nelson, B., Hay, J.A., Monod, M., Miscouridou, X., Coupland, H., Sonabend, R., Vollmer, M., Gandy, A., Prete, C.A., Nascimento, V.H., Suchard, M.A., Bowden, T.A., Pond, S.L.K., Wu, C.-H., Ratmann, O., Ferguson, N.M., Dye, C., Loman, N.J., Lemey, P., Rambaut, A., Fraiji, N.A., Carvalho, M. do P.S.S., Pybus, O.G., Flaxman, S., Bhatt, S., Sabino, E.C., 2021. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science 372, 815-821. https://doi.org/10.1126/science.abh2644

Fehr, A.R., Perlman, S., 2015. Coronaviruses: An Overview of Their Replication and Pathogenesis. Methods Mol. Biol. Clifton NJ 1282, 1–23. https://doi.org/10.1007/978-1-4939-2438-7_1

Ferrareze, P.A.G., Franceschi, V.B., Mayer, A. de M., Caldana, G.D., Zimerman, R.A., Thompson, C.E., 2021. E484K as an innovative phylogenetic event for viral evolution: Genomic analysis of the E484K spike mutation in SARS-CoV-2 lineages from Brazil. Infect. Genet. Evol. 93, 104941. https://doi.org/10.1016/j.meegid.2021.104941

Fourati, S., Decousser, J.-W., Khouider, S., N'Debi, M., Demontant, V., Trawinski, E., Gourgeon, A., Gangloff, C., Destras, G., Bal, A., Josset, L., Soulier, A., Costa, Y., Gricourt, G., Lina, B., Lepeule, R., Pawlotsky, J.-M., Rodriguez, C., 2021. Novel SARS-CoV-2 Variant Derived from Clade 19B, France. Emerg. Infect. Dis. 27. https://doi.org/10.3201/eid2705.210324

Franceschi, V.B., Ferrareze, P.A.G., Zimerman, R.A., Cybis, G.B., Thompson, C.E., 2021. Mutation hotspots, geographical and temporal distribution of SARS-CoV-2 lineages in Brazil, February 2020 to February 2021: insights and limitations from uneven sequencing efforts. medRxiv 2021.03.08.21253152. https://doi.org/10.1101/2021.03.08.21253152

Frappier, V., Najmanovich, R.J., 2014. A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. PLoS Comput. Biol. 10, e1003569. https://doi.org/10.1371/journal.pcbi.1003569

Gibbs, M.J., Armstrong, J.S., Gibbs, A.J., 2000. Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16, 573–582. https://doi.org/10.1093/bioinformatics/16.7.573

Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M., McCammon, J.A., Caves, L.S.D., 2006. Bio3d: an R package for the comparative analysis of protein structures. Bioinforma. Oxf. Engl. 22, 2695–2696. https://doi.org/10.1093/bioinformatics/btl461

Greaney, A.J., Starr, T.N., Gilchuk, P., Zost, S.J., Binshtein, E., Loes, A.N., Hilton, S.K., Huddleston, J., Eguia, R., Crawford, K.H.D., Dingens, A.S., Nargi, R.S., Sutton, R.E., Suryadevara, N., Rothlauf, P.W., Liu, Z., Whelan, S.P.J., Carnahan, R.H., Crowe, J.E., Bloom, J.D., 2020. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. Cell Host Microbe. https://doi.org/10.1016/j.chom.2020.11.007

Groves, D.C., Rowland-Jones, S.L., Angyal, A., 2021. The D614G mutations in the SARS-CoV-2 spike protein: Implications for viral infectivity, disease severity and vaccine design. Biochem. Biophys. Res. Commun., COVID-19 538, 104–107. https://doi.org/10.1016/j.bbrc.2020.10.109

Gu, H., Chen, Q., Yang, G., He, L., Fan, H., Deng, Y.-Q., Wang, Y., Teng, Y., Zhao, Z., Cui, Y., Li, Yuchang, Li, X.-F., Li, J., Zhang, N.-N., Yang, Xiaolan, Chen, S., Guo, Y., Zhao, G., Wang, X., Luo, D.-Y., Wang, H., Yang, Xiao, Li, Yan, Han, G., He, Y., Zhou, X., Geng, S., Sheng, X., Jiang, S., Sun, S., Qin, C.-F., Zhou, Y., 2020. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. Science 369, 1603–1607. https://doi.org/10.1126/science.abc4730

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst. Biol. 59, 307–321. https://doi.org/10.1093/sysbio/syq010

Guo, K., Barrett, B.S., Mickens, K.L., Hasenkrug, K.J., Santiago, M.L., 2021. Interferon Resistance of Emerging SARS-CoV-2 Variants. bioRxiv 2021.03.20.436257. https://doi: 10.1101/2021.03.20.436257

Hark Gan, H., Twaddle, A., Marchand, B., Gunsalus, K.C., 2021. Structural modeling of the SARS-CoV-2 Spike/human ACE2 complex interface can identify high-affinity variants associated with increased transmissibility. J. Mol. Biol. 14, 433(15):167051. https://doi: 10.1016/j.jmb.2021.167051

Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S.J., Robertson, D.L., 2021. SARS-CoV-2 variants, spike mutations and immune escape. Nat. Rev. Microbiol. 1–16. https://doi.org/10.1038/s41579-021-00573-0

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S., 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35, 518–522. https://doi.org/10.1093/molbev/msx281

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M.A., Drosten, C., Pöhlmann, S., 2020. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell 181, 271-280.e8.

https://doi.org/10.1016/j.cell.2020.02.052
Holmes, E.C., Worobey, M., Rambaut, A., 1999. Phylogenetic evidence for recombination in dengue virus. Mol. Biol. Evol. 16, 405–409. https://doi.org/10.1093/oxfordjournals.molbev.a026121
Homeyer, N., Gohlke, H., 2012. Free Energy Calculations by the Molecular Mechanics Poisson−Boltzmann Surface Area Method. Mol. Inform. 31, 114–122. https://doi.org/10.1002/minf.201100135
Hoover, W.G., 1985. Canonical dynamics: Equilibrium phase-space distributions. Phys. Rev. A 31, 1695–1697. https://doi.org/10.1103/PhysRevA.31.1695
Humphrey, W., Dalke, A., Schulten, K., 1996. VMD: Visual molecular dynamics. J. Mol. Graph. 14, 33–38. https://doi.org/10.1016/0263-7855(96)00018-5
Jackson, B., Rambaut, A., Pybus, O., Robertson, D.L., Connor, T., Loman, N., The COVID-19 Genomics UK (COG-UK) consortium, 2021. Recombinant SARS-CoV-2 genomes involving lineage B.1.1.7 in the [WWW Document]. Virological. URL https://virological.org/t/recombinant-sars-cov-2-genomes-involving-lineage-b-1-1-7-in-the-uk/658 (accessed 3.30.21).
Jackson, C.B., Zhang, L., Farzan, M., Choe, H., 2021. Functional importance of the D614G mutation in the SARS-CoV-2 spike protein. Biochem. Biophys. Res. Commun., COVID-19 538, 108–115. https://doi.org/10.1016/j.bbrc.2020.11.026
Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L., 1983. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 79, 926–935. https://doi.org/10.1063/1.445869
Katoh, K., Rozewicki, J., Yamada, K.D., 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20, 1160–1166. https://doi.org/10.1093/bib/bbx108
Kemp, S.A., Collier, D.A., Datir, R.P., Ferreira, I.A.T.M., Gayed, S., Jahun, A., Hosmillo, M., Rees-Spear, C., Mlcochova, P., Lumb, I.U., Roberts, D.J., Chandra, A., Temperton, N., Sharrocks, K., Blane, E., Modis, Y., Leigh, K., Briggs, J., van Gils, M., Smith, K.G.C., Bradley, J.R., Smith, C., Doffinger, R., Ceron-Gutierrez, L., Barcenas-Morales, G., Pollock, D.D., Goldstein, R.A., Smielewska, A., Skittrall, J.P., Gouliouris, T., Goodfellow, I.G., Gkrania-Klotsas, E., Illingworth, C.J.R., McCoy, L.E., Gupta, R.K., 2021. SARS-CoV-2 evolution during treatment of chronic infection. Nature 1–10. https://doi.org/10.1038/s41586-021-03291-y
Kim, Y.-M., Shin, E.-C., 2021. Type I and III interferon responses in SARS-CoV-2 infection. Experimental & Molecular Medicine 53, 750–760. https://doi.org/10.1038/s12276-021-00592-0
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., Angyal, A., Brown, R.L., Carrilero, L., Green, L.R., Groves, D.C., Johnson, K.J., Keeley, A.J., Lindsey, B.B., Parsons, P.J., Raza, M., Rowland-Jones, S., Smith, N., Tucker, R.M., Wang, D., Wyles, M.D., McDanal, C., Perez, L.G., Tang, H., Moon-Walker, A., Whelan, S.P., LaBranche, C.C., Saphire, E.O., Montefiori, D.C., 2020. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. https://doi.org/10.1016/j.cell.2020.06.043
Kosakovsky Pond, S.L., Frost, S.D.W., 2005. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. Mol. Biol. Evol. 22, 1208–1222. https://doi.org/10.1093/molbev/msi105
Kumari, R., Kumar, R., Lynn, A., 2014. g_mmpbsa—A GROMACS Tool for High-Throughput MM-PBSA Calculations. J. Chem. Inf. Model. 54, 1951–1962. https://doi.org/10.1021/ci500020m

Laimer, J., Hiebl-Flach, J., Lengauer, D., Lackner, P., 2016. MAESTROweb: a web server for structure-based protein stability prediction. Bioinforma. Oxf. Engl. 32, 1414–1416. https://doi.org/10.1093/bioinformatics/btv769

Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26, 283–291. https://doi.org/10.1107/S0021889892009944

Liu, D.X., Fung, T.S., Chong, K.K.-L., Shukla, A., Hilgenfeld, R., 2014. Accessory proteins of SARS-CoV and other coronaviruses. Antiviral Res. 109, 97–109. https://doi.org/10.1016/j.antiviral.2014.06.013

Martin, D., Rybicki, E., 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16, 562–563. https://doi.org/10.1093/bioinformatics/16.6.562

Martin, D.P., Murrell, B., Golden, M., Khoosal, A., Muhire, B., 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 1. https://doi.org/10.1093/ve/vev003

Martin, D.P., Posada, D., Crandall, K.A., Williamson, C., 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res. Hum. Retroviruses 21, 98–102. https://doi.org/10.1089/aid.2005.21.98

Martin, D.P., Weaver, S., Tegally, H., San, E.J., Shank, S.D., Wilkinson, E., Giandhari, J., Naidoo, S., Pillay, Y., Singh, L., Lessells, R.J., NGS-SA, Uk (cog-Uk), C.-19 G., Gupta, R.K., Wertheim, J.O., Nekturenko, A., Murrell, B., Harkins, G.W., Lemey, P., MacLean, O.A., Robertson, D.L., Oliveira, T. de, Pond, S.L.K., 2021. The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. medRxiv 2021.02.23.21252268. https://doi.org/10.1101/2021.02.23.21252268

Meng, B., Kemp, S.A., Papa, G., Datir, R., Ferreira, I.A.T.M., Marelli, S., Harvey, W.T., Lytras, S., Mohamed, A., Gallo, G., Thakur, N., Collier, D.A., Mlcochova, P., Duncan, L.M., Carabelli, A.M., Kenyon, J.C., Lever, A.M., De Marco, A., Saliba, C., Culap, K., Cameroni, E., Matheson, N.J., Piccoli, L., Corti, D., James, L.C., Robertson, D.L., Bailey, D., Gupta, R.K., 2021. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. Cell Reports 35, 109292. https://doi.org/10.1016/j.celrep.2021.109292

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S.L., Scheffler, K., 2013. FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. Mol. Biol. Evol. 30, 1196–1205. https://doi.org/10.1093/molbev/mst030

Nelson, G., Buzko, O., Spilman, P., Niazi, K., Rabizadeh, S., Soon-Shiong, P., 2021. Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. bioRxiv 2021.01.13.426558. https://doi.org/10.1101/2021.01.13.426558

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300

Nosé, S., 1984. A unified formulation of the constant temperature molecular dynamics methods. J. Chem. Phys. 81, 511–519. https://doi.org/10.1063/1.447334

Padidam, M., Sawyer, S., Fauquet, C.M., 1999. Possible Emergence of New Geminiviruses by Frequent Recombination. Virology 265, 218–225. https://doi.org/10.1006/viro.1999.0056

Parrinello, M., Rahman, A., 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. J. Appl. Phys. 52, 7182–7190. https://doi.org/10.1063/1.328693

Pereson, M.J., Mojsiejczuk, L., Martínez, A.P., Flichman, D.M., Garcia, G.H., Di Lello, F.A., 2020. Phylogenetic analysis of SARS-CoV-2 in the first few months since its emergence. Journal of Medical Virology 93, 1722–1731.

https://doi.org/10.1002/jmv.26545
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., Plante, K.S., Weaver, S.C., Shi, P.-Y., 2020. Spike mutation D614G alters SARS-CoV-2 fitness. Nature 1–6. https://doi.org/10.1038/s41586-020-2895-3
Poon, A.F.Y., Lewis, F.I., Pond, S.L.K., Frost, S.D.W., 2007. An Evolutionary-Network Model Reveals Stratified Interactions in the V3 Loop of the HIV-1 Envelope. PLOS Comput. Biol. 3, e231. https://doi.org/10.1371/journal.pcbi.0030231
Posada, D., Crandall, K.A., 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Natl. Acad. Sci. 98, 13757–13762.
Pucci, F., Rooman, M., 2021. Prediction and Evolution of the Molecular Fitness of SARS-CoV-2 Variants: Introducing SpikePro. Viruses 13, 935. https://doi.org/10.3390/v13050935
Qu, X.-X., Hao, P., Song, X.-J., Jiang, S.-M., Liu, Y.-X., Wang, P.-G., Rao, X., Song, H.-D., Wang, S.-Y., Zuo, Y., Zheng, A.-H., Luo, M., Wang, H.-L., Deng, F., Wang, H.-Z., Hu, Z.-H., Ding, M.-X., Zhao, G.-P., Deng, H.-K., 2005. Identification of Two Critical Amino Acid Residues of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein for Its Variation in Zoonotic Tropism Transition via a Double Substitution Strategy. Journal of Biological Chemistry 280, 29588–29595. https://doi.org/10.1074/jbc.m500662200
Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D., Volz, E., 2020. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations [WWW Document]. Virological. URL https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (accessed 1.4.21).
Rathnasinghe, R., Jangra, S., Cupic, A., Martínez-Romero, C., Mulder, L.C.F., Kehrer, T., Yildiz, S., Choi, A., Mena, I., De Vrieze, J., Aslam, S., Stadlbauer, D., Meekins, D.A., McDowell, C.D., Balaraman, V., Richt, J.A., De Geest, B.G., Miorin, L., Krammer, F., Simon, V., García-Sastre, A., Schotsaert, M., 2021. The N501Y mutation in SARS-CoV-2 spike leads to morbidity in obese and aged mice and is neutralized by convalescent and post-vaccination human sera. Cold Spring Harbor Laboratory.
Rodrigues, C.H., Pires, D.E., Ascher, D.B., 2018. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res. 46, W350–W355. https://doi.org/10.1093/nar/gky300
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, W382–W388. https://doi.org/10.1093/nar/gki387
Smith, J.M., 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. 34, 126–129. https://doi.org/10.1007/BF00182389
Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H.D., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C., King, N.P., Veesler, D., Bloom, J.D., 2020. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell 182, 1295-1310.e20. https://doi.org/10.1016/j.cell.2020.08.012
Studer, R.A., Christin, P.-A., Williams, M.A., Orengo, C.A., 2014. Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. Proceedings of the National Academy of Sciences 111, 2223–2228. https://doi.org/10.1073/pnas.1310811111
Tang, J.W., Tambyah, P.A., Hui, D.S., 2021. Emergence of a new SARS-CoV-2 variant in the UK. J. Infect. 82, e27–e28. https://doi.org/10.1016/j.jinf.2020.12.024
Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E.J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza, S., Allam, M.,

Ismail, A., Mohale, T., Glass, A.J., Engelbrecht, S., Van Zyl, G., Preiser, W., Petruccione, F., Sigal, A., Hardie, D., Marais, G., Hsiao, N., Korsman, S., Davies, M.-A., Tyers, L., Mudau, I., York, D., Maslo, C., Goedhals, D., Abrahams, S., Laguda-Akingba, O., Alisoltani-Dehkordi, A., Godzik, A., Wibmer, C.K., Sewell, B.T., Lourenço, J., Alcantara, L.C.J., Kosakovsky Pond, S.L., Weaver, S., Martin, D., Lessells, R.J., Bhiman, J.N., Williamson, C., de Oliveira, T., 2021. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature 592, 438–443. https://doi.org/10.1038/s41586-021-03402-9

Van Egeren, D., Novokhodko, A., Stoddard, M., Tran, U., Zetter, B., Rogers, M., Pentelute, B.L., Carlson, J.M., Hixon, M., Joseph-McCarthy, D., Chakravarty, A., 2021. Risk of rapid evolutionary escape from biomedical interventions targeting SARS-CoV-2 spike protein. PLOS ONE 16, e0250780. https://doi.org/10.1371/journal.pone.0250780

Walls, A.C., Park, Y.-J., Tortorici, M.A., Wall, A., McGuire, A.T., Veesler, D., 2020. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 181, 281-292.e6. https://doi.org/10.1016/j.cell.2020.02.058

Wang, R., Chen, J., Gao, K., Wei, G.-W., 2021. Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries. Genomics 113, 2158–2170. https://doi.org/10.1016/j.ygeno.2021.05.006

Washington, N.L., Gangavarapu, K., Zeller, M., Bolze, A., Cirulli, E.T., Schiabor Barrett, K.M., Larsen, B.B., Anderson, C., White, S., Cassens, T., Jacobs, S., Levan, G., Nguyen, J., Ramirez, J.M., III, Rivera-Garcia, C., Sandoval, E., Wang, X., Wong, D., Spencer, E., Robles-Sikisaka, R., Kurzban, E., Hughes, L.D., Deng, X., Wang, C., Servellita, V., Valentine, H., De Hoff, P., Seaver, P., Sathe, S., Gietzen, K., Sickler, B., Antico, J., Hoon, K., Liu, J., Harding, A., Bakhtar, O., Basler, T., Austin, B., MacCannell, D., Isaksson, M., Febbo, P.G., Becker, D., Laurent, M., McDonald, E., Yeo, G.W., Knight, R., Laurent, L.C., de Feo, E., Worobey, M., Chiu, C.Y., Suchard, M.A., Lu, J.T., Lee, W., Andersen, K.G., 2021. Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. Cell 184, 2587-2594.e7. https://doi.org/10.1016/j.cell.2021.03.052

Webb, B., Sali, A., 2016. Comparative Protein Structure Modeling Using MODELLER. Curr. Protoc. Bioinforma. 54, 5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3

Weiller, G.F., 1998. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol. Biol. Evol. 15, 326–335. https://doi.org/10.1093/oxfordjournals.molbev.a025929

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., Robbiani, D.F., Rice, C.M., Nussenzweig, M.C., Hatziioannou, T., Bieniasz, P.D., 2020. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. eLife 9, e61312. https://doi.org/10.7554/eLife.61312

Weissman, D., Alameh, M.-G., Silva, T. de, 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.C., Tam, Y., Shaw, P.A., Lewis, M.G., Boesler, C., Şahin, U., Acharya, P., Haynes, B.F., Korber, B., Montefiori, D.C., 2021. D614G Spike Mutation Increases SARS CoV-2 Susceptibility to Neutralization. Cell Host Microbe 29, 23-31.e4. https://doi.org/10.1016/j.chom.2020.11.012

Winger, A., Caspari, T., 2021. The Spike of Concern—The Novel Variants of SARS-CoV-2. Viruses 13. https://doi.org/10.3390/v13061002

Yuan, M., Wu, N.C., Zhu, X., Lee, C.-C.D., So, R.T.Y., Lv, H., Mok, C.K.P., Wilson, I.A., 2020. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science 368, 630–633. https://doi.org/10.1126/science.abb7269

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., Dudkina, N.V., Shen, K., Luban, J., 2020. Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. Cell 183, 739-751.e8. https://doi.org/10.1016/j.cell.2020.09.032

Supplementary figures

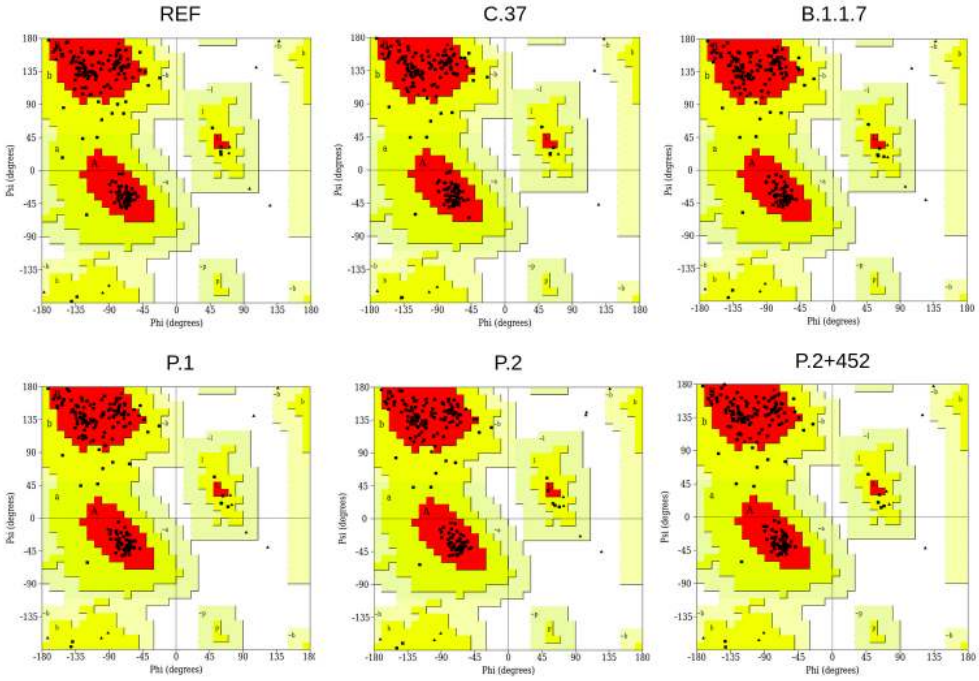

Figure S1. Ramachandran plots indicating the quality of the modeled structures of the spike RBD - hACE2 protein complexes for the reference and lineages P.1, P.2, C.37, B.1.1.7, and P.2+452.

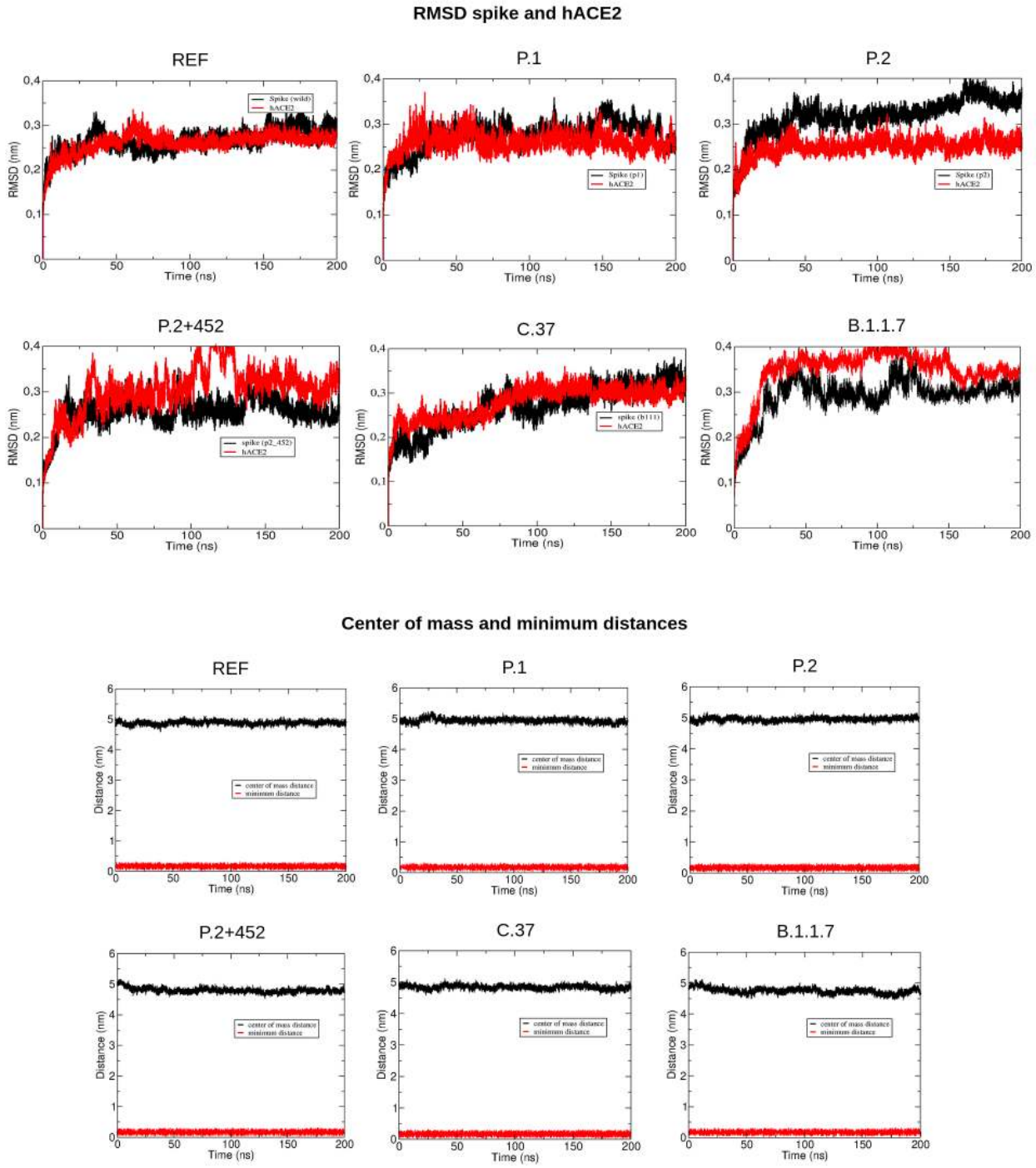

Figure S2. Root-mean-square deviation (RMSD), center of mass and minimum distances for the spike RBD - ACE2 protein complexes related to the reference (REF) and lineages P.1, P.2, P.2+452, C.37, and B.1.1.7.

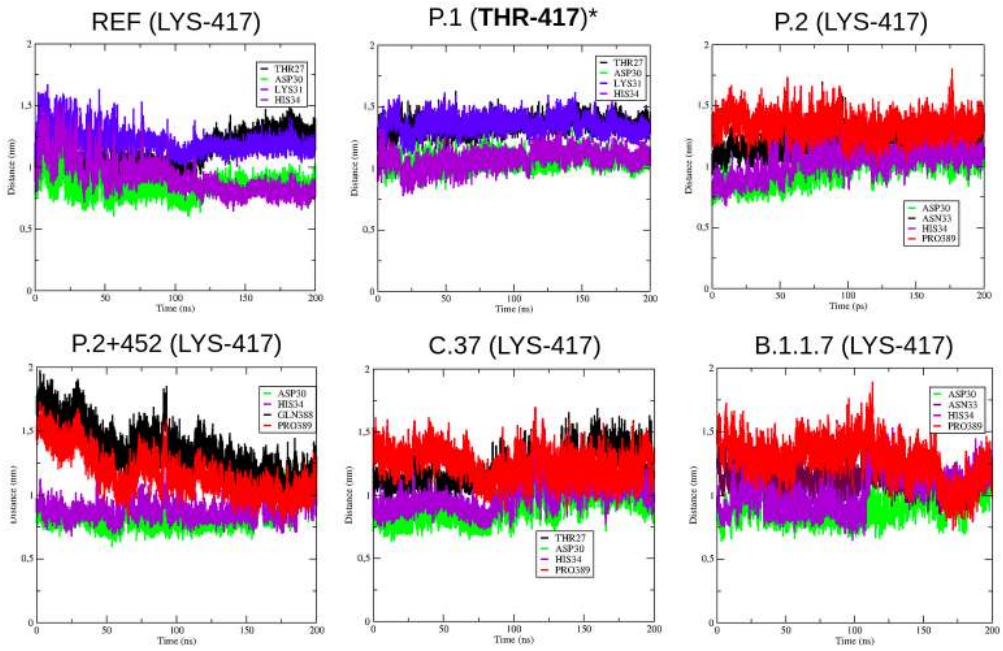

Figure S3. Structural distance measurements to the residue 417 in the reference RBD-ACE2 complex and for the lineages C.37, B.1.1.7, P.1, P.2, and P.2+452.

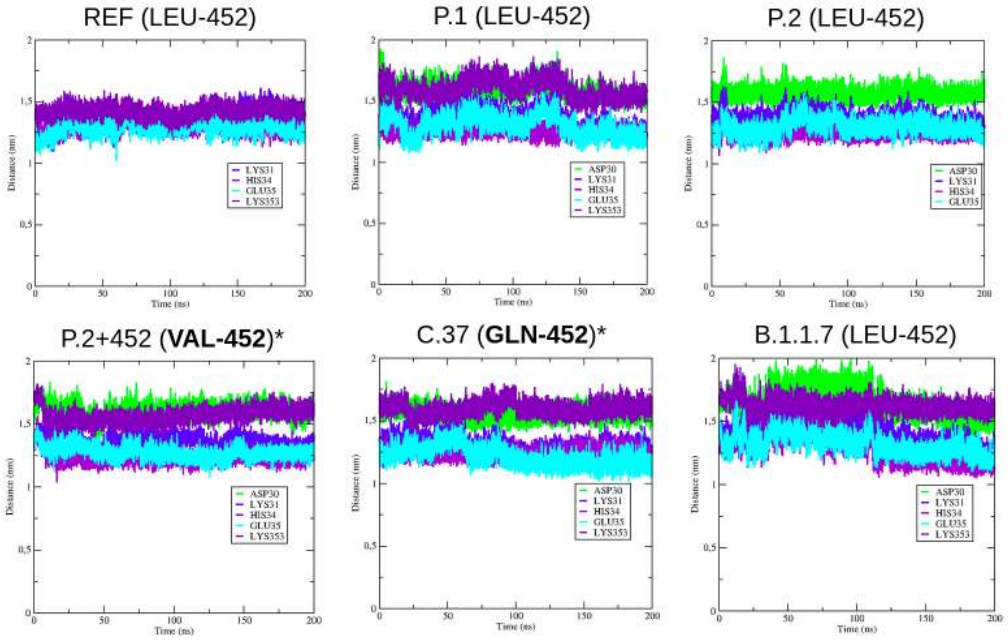

Figure S4. Structural distance measurements to the residue 452 in the reference RBD-ACE2 complex and for the lineages C.37, B.1.1.7, P.1, P.2, and P.2+452.

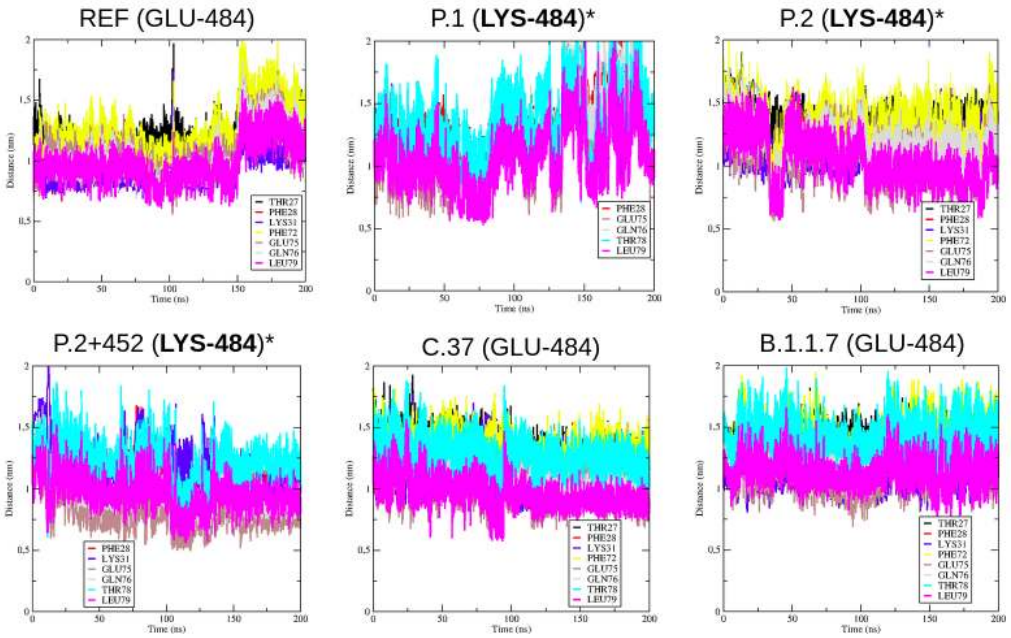

Figure S5. Structural distance measurements to the residue 484 in the reference RBD-ACE2 complex and for the lineages C.37, B.1.1.7, P.1, P.2, and P.2+452.

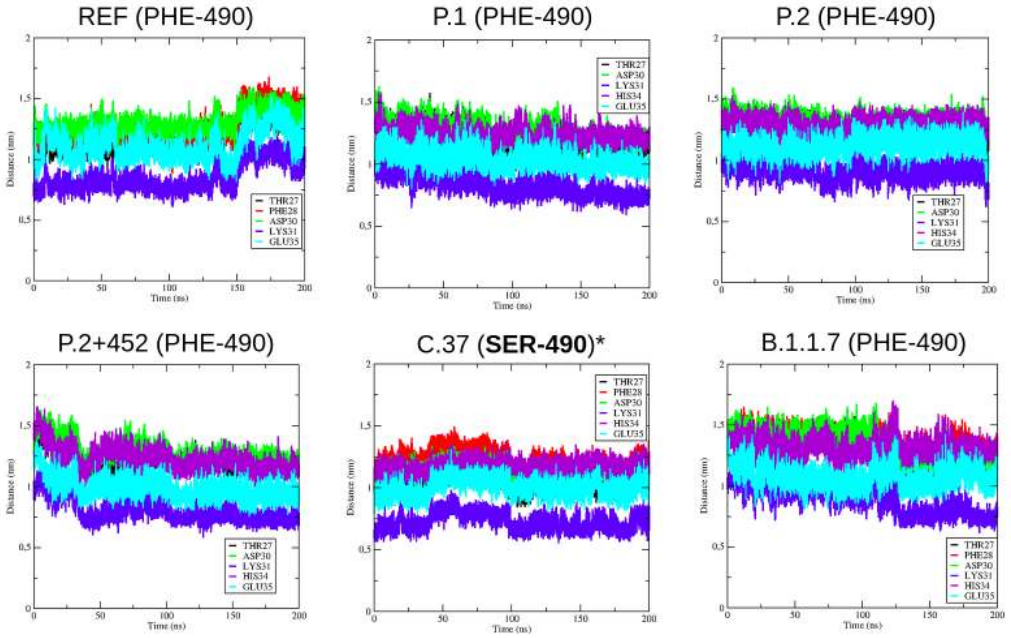

Figure S6. Structural distance measurements to the residue 490 in the reference RBD-ACE2 complex and for the lineages C.37, B.1.1.7, P.1, P.2, and P.2+452.

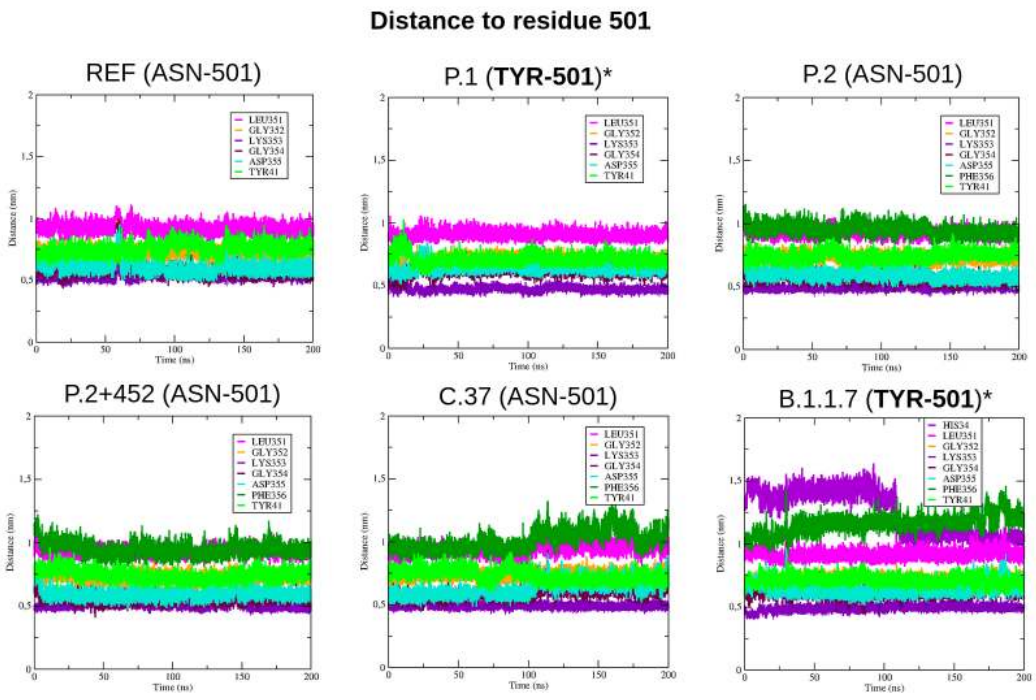

Figure S7. Structural distance measurements to the residue 501 in the reference RBD-ACE2 complex and for the lineages C.37, B.1.1.7, P.1, P.2, and P.2+452.

Supplementary files

Supplementary File 1. Acknowledgement list of the 11,078 GISAID genomes used in this study.

Supplementary File 2. Complete table of positively selected sites identified by FUBAR, FEL, and SLAC methods.

Supplementary File 3. Table of sites under purifying selection according to the FEL and SLAC methods.

Supplementary File 4. Complete table of sites not conditionally independent identified by the Bayesian Graphical Model (BGM).

Supplementary File 5. Table of evaluated parameters from PROCHECK/VERIFY3D for the spike RBD models selection.