UNIVERSITAT ROVIRA i VIRGILI

BEYOND SINGLE MUTATIONS: EXPLORING SIMULTANEOUS MUTATIONS IN SARS-CoV-2 M-pro AND Spike PROTEINS

FINAL DEGREE PROJECT

Degree in Biochemistry and Molecular Biology

Faculty of Chemistry

Rovira and Virgili University

Year 2022-2023

MENTOR : SANTIAGO GARCÍA VALLVÉ

ALBERTO CONSTANTINO PUSCASU

INDEX

ABSTRACT

Understanding the genetic variability and mutational patterns of SARS-CoV-2 is crucial for effective and control and prevention of infectious diseases such as COVID-19. In this study we examined a dataset comprising 922,474 mutations in the M-pro protein and 44,278,219 mutations in the Spike protein of SARS-CoV-2. Our research aimed on finding simultaneous mutations on those proteins and trying to understand their biological reason.

Our research shows that both the M-pro and Spike proteins have simultaneous mutations. We discovered that these simultaneous mutations are not directly implied on maintaining the structure of the protein. Furthermore, we determined that simultaneous mutations are not correlated with alterations in the secondary RNA structure of SARS-CoV-2.

This study sheds lights on the significance of simultaneous mutations and their relationship with Variants of Concern (VOCs). Simultaneous mutations in the Spike protein have been grouped into clusters that define the signature mutations of different VOCs.

INTRODUCTION

It is well known that the pandemic associated with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been one of the latest and most lethal pandemics our society has suffered. However, this has not been the first outbreak of this type of virus as the severe acute respiratory syndrome (SARS) and the Middle East respiratory syndrome (MERS) began to emerge in 2002 and 2012, respectively. The disease caused by SARS-CoV-2, first identified in Wuhan City, Hubei Province, China, on 12 December 2019, was named COVID-19 [1].

Coronaviruses (CoVs) belong to the family of *Coronaviridae* (subfamily *Coronavirinae*). These viruses can infect a broad range of hosts, producing symptoms and diseases ranging from the common cold to severe and fatal illnesses such as COVID-19. Six CoVs were known to infect humans until 2020 when SARS-CoV-2 was discovered. Coronaviruses are enveloped positive-sense RNA viruses characterized by club-like spikes that project from their surface and an unusual large RNA genome of s ~30 kb [1,2]. This genome is non-segmented and contains a 5' cap structure along with a 3' poly (A) tail. It also presents a highly 5' untranslated region (5'-UTR) that plays a key role in the regulation of RNA replication and translation. The SARS-CoV-2 genome contains 14 Open Reading Frames (ORFs), preceded by transcriptional regulatory sequences [3]. The 5' two thirds of the genome consist of the replicase genes for large polyproteins, PP1a and PP1ab inside the two main transcriptional units ORF1a and ORF1ab. ORF1a allows for the synthesis of a polyprotein that is later cleaved into the non-structural proteins NSP1 to NSP11 by proteolytic cleavages. One frameshifting (-1) event is necessary for the translation of the second ORF1ab, and the produced polyprotein is further processed by proteolytic cleavages to produce four more non-structural proteins, NSP12 to NSP16, which form the complex replicase machinery [3,4]. The remaining third of the genome is encoding structural proteins like Spike (S), Envelope (E), Membrane (M) and Nucleocapsid (N) which are components of the mature virus and play a crucial role in viral structure integrity and entry into the host. This part of the genome also contains nine putative ORFs for accessory proteins that are translated into sub-genomic RNA (sgRNA) that start with 5'UTR leader and terminate with the 3'UTR end of the full-length gRNA (Figure 1) [3,5,6].

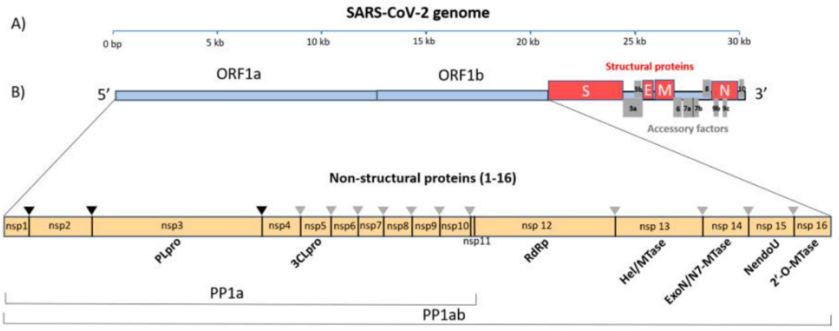

Figure 1: SARS-CoV-2 polycistronic genome. A) Genome of SARS-CoV-2 organized in individual ORFs. B) Polyprotein 1ab embeds 16 non-structural proteins, the black and grey triangles indicate the cleavage sites of the protease PLpro and 3CLpro (M-pro), respectively. Extracted from [3].

Subgenomic RNAs (sgRNA) are synthesized by the viral RNA-dependent polymerase RdRp (NSP12). It uses Transcription regulatory Sequences (TRSs) located at the 5' end (TRS-B) and 5' leader (TRS-L) of each subgenomic coding sequence (see TRS-B and TRS-L on Figure 2). The RNA polymerase RdRp pauses on TRS-B sequences and shifts the template to the TRS-L through discontinuous transcription when it undergoes negative strand synthesis from the 3' end of the genomic RNA [4].

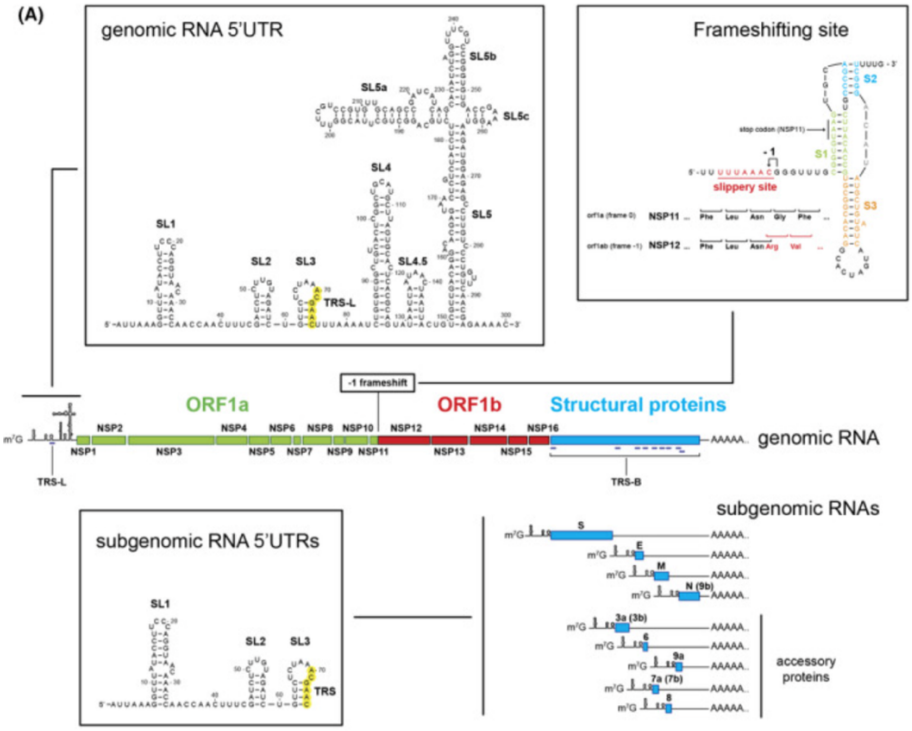

Figure 2: Genomic and subgenomic RNA transcripts. The Programmed -1 Frameshift Stimulation Element and the genomic and subgenomic RNAs' 5'UTR secondary structures are indicated in boxes. The 5'UTRs display the TRSL and TRS nucleotides in green. Adapted from [4]

This process enables the synthesis of negative strand templates that are then utilised for the synthesis of positive stranded sgRNAs [4,6]. The full-length (+) gRNA serves both to synthetise (-) gRNA needed for the genome replication and also to synthetise (-) sgRNA which are consequently used to generate (+) subgenomic mRNA that will translate into structural and accessory proteins (Figure 3) [6].

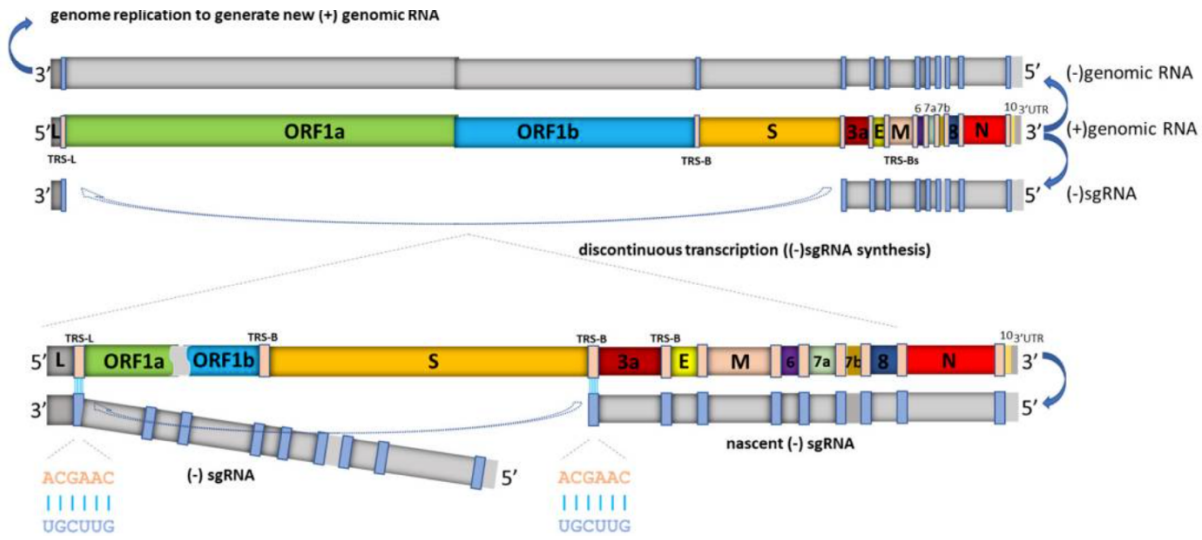

Figure 3: schematic representation of SARS-CoV-2 subgenomic discontinuous transcription. The TRS-B and TRS-L are shown as nucleotides in both 3' end and start of the sgRNA genome part. Extracted from [6]

Nsp5, also known as 3CLpro or Main protease (M-pro) is a 33.8kDa cysteine protease, which processes the two replicase polyproteins PP1a of 486kDa and PP1ab of 790kDa. M-pro has 3 domains: domain 1 containing residues from 8 to 101, domain 2 containing residues from 101 to 184 and domain 3 containing residues from 201 to 303. The first two domains are composed of antiparallel –barrel structures and are the catalytic domains. The third domain is composed of five –helices and is responsible for the enzyme dimerization. It also has a long loop between domain 2 and 3 [7]. The active site of M-pro is located between the gaps of domains 1 and 2 and has a CysHis catalytic dyad (Cys145 and His41). The sulphur of the cysteine serves as a nucleophile and the imidazole ring of the histidine as a general base [7,8]. The catalytic mechanism of M-pro protein on the hydrolysis of amide substrate is shown on Figure 4.

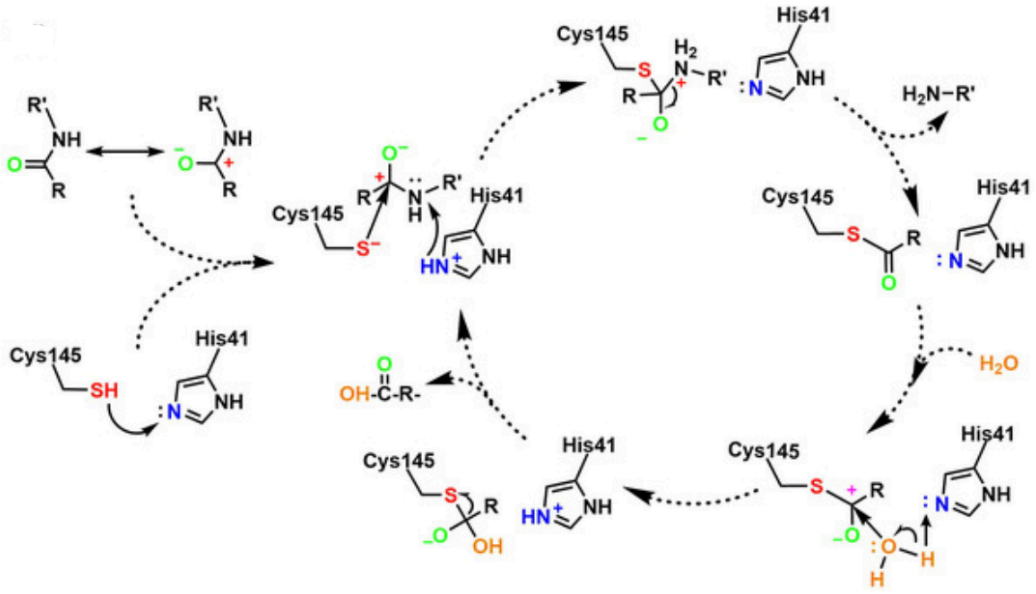

Figure 4: the catalytic mechanism of M-pro protein on the hydrolysis of amide substrate. Adapted from [8]

M-pro is first cleaved from polyproteins and once it is a mature enzyme it cleaves downstream nsps at 11 sites to release from nsp4 to nsp16. These cleavages are done because M-pro can break a bond between glutamine at position P1 and a small amino acid such as serine alanine or glycine at position P2. As M-pro is a protein that cleaves directly de nsp proteins, it is a crucial factor in the virus life cycle, and it is an attractive target for drug development against the disease [7]. The crystal structure of M-pro has been reported (PDB ID: 6Y2E) and is shown on Figure 5A.

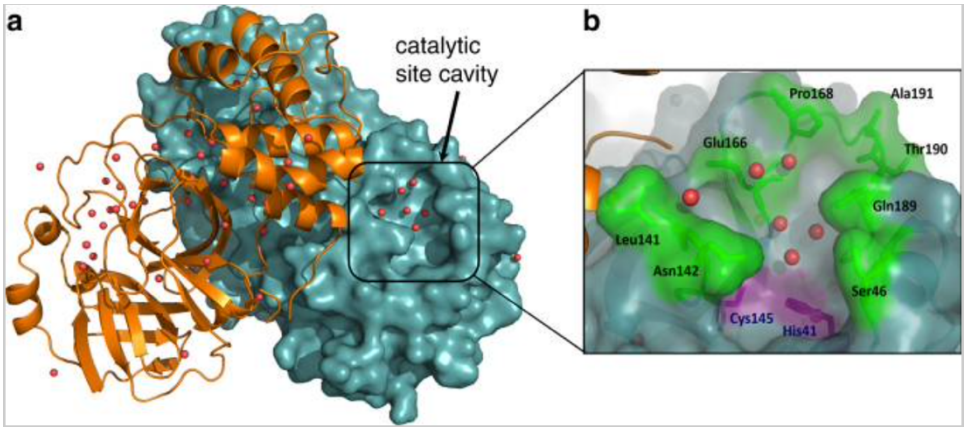

Figure 5: Three-dimensional structure of 3CLpro from SARS-CoV-2. A) One monomer of the dimer is shown as an orange cartoon and the other monomer is shown as a teal surface with the catalytic site cavity highlighted with water molecules shown as red spheres. B) A closeup view of the catalytic site cavity in which the catalytic residues are highlighted in purple, the residues that flank the cavity in green and water molecules are red spheres. Extracted from [9]

M-pro protein is an interesting protein because it has one of the lowest mutation rates. Having collected 5,340,569 high-coverage SARS-CoV-2 genomes, available until June 27, 2022 from the Global Initiative on Sharing Avian Influenza Data (GISAID) database and compared with the reference SARS-CoV-2 genome (NC_045512.2) isolated in December 2019 from Wuhan-Hu-1 it was determined that the M-pro gene (nsp5) is one with the lowest nonsynonymous mutation rates (Figure 6) [10].

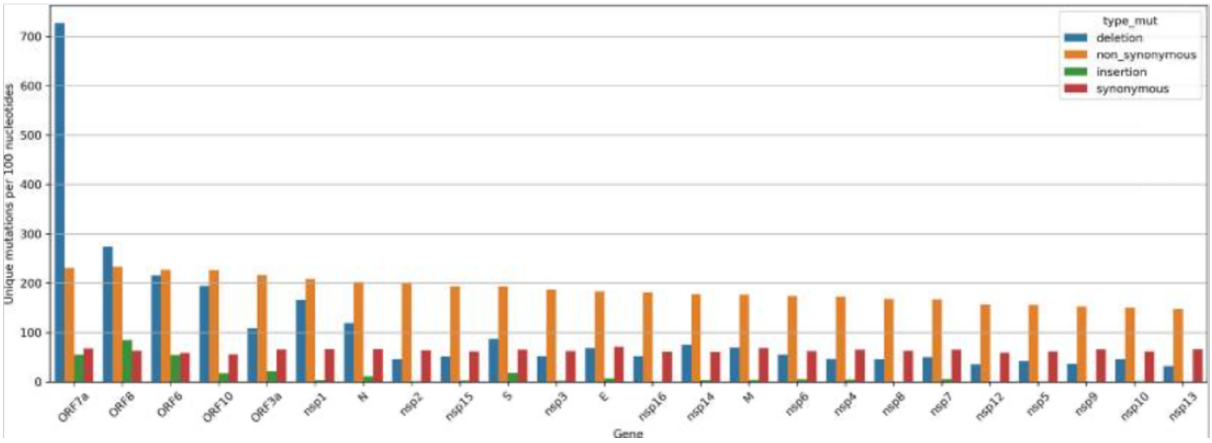

Figure 6: Unique mutations per 100 nucleotides in the SARS-CoV-2 genes. Deletions, nonsynonymous mutations, insertions, and synonymous mutations are shown in blue, yellow, green, and red, respectively. Extracted from [10]

Spike (S) proteins are structural glycoproteins of 180-200kDa that are involved in the host-virus infection. They fuse with the membrane of the host cell and allows the virus to enter it. The Spike proteins are targeted proteins for different drug design and neutralization antibodies because its superficial localization. Spike proteins consist of two subunits and a signal peptide (amino acids 1-13). First subunit S1 (14-685 residues) binds with the host cellular receptors and second subunit S2 (686-1273 residues) facilitates the membrane fusion of the virus with the host cell [11]. The structure of Spike proteins is shown in Figure 6. The S1 subunit is formed by a N-terminal domain (NTD) and a receptor-binding domain (RBD). The S2 subunit is formed by the fusion peptide (FP), heptapeptide repeat sequences 1 and 2 (HR1 and HR2), transmembrane domain (TM) and cytoplasm domain (CT). Spike protein trimers visually form a characteristic bulbous, crown-like halo surrounding the viral particle [11]. The trimers can be found in open or close conformations (Figure 7) depending on several factors, including the binding of the receptor-binding domain (RBD) to the host cell receptor, the proteolytic cleavage of the Spike protein by host cell proteases, and the interaction with other viral proteins. In the closed state, the RBD is inaccessible

to the host cell receptor, whereas in the open state, the RBD is exposed and able to bind to the host cell receptor. The conformational changes of the Spike protein are critical for the virus's ability to infect host cells and evade host immune responses [11].

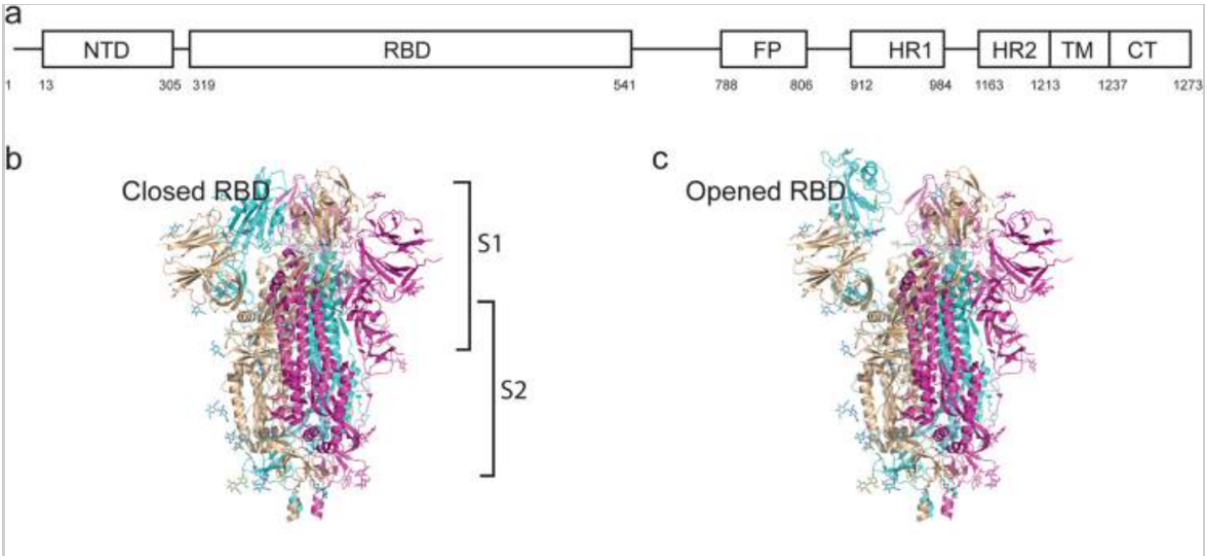

Figure 7: Structure of the SARS-CoV-2. A) Schematic representation of the SARS-CoV-2 Spike protein. B-C) The Spike protein RBD closed and opened conformation. Adapted from [11]

The SARS-CoV-2 Spike protein binds to the host cell by recognizing the receptor ACE2, an ACE homologue that converts angiotensin I to angiotensin 1-9 [11]. The Spike protein I assembled as a homotrimer and is inserted multiple times into the membrane of the virion giving it its crown-like appearance [12]. The Spike protein binds to ACE2 through the RBD region of the S1 subunit and the process of viral fusion to the host cell is mediated by the cleavage of S1 and S2 subunits by host proteases [11–13]. Both subunits exist in a noncovalent form until viral fusion occurs. The cleavage site of the Spike protein is done in specific furin cleavage sites and the SARS-CoV-2 S has multiple furin cleavage sites, which increases the probability of being cleaved by furin-like proteases and thereby enhances its efficacy [11,12]. TMPRSS2 and trypsin are some human proteases that have been proven to cleave Spike proteins. After the cleavage of the Spike protein, the FP domain of SARS-CoV-2 S2 subunit is exposed and triggers viral fusion [11–13]. FP domain interacts with the host cell membrane and brings the viral and host cell membrane into proximity. This process provokes a structural change in S2 subunit into a post-fusion state forming a trimeric coiled-coil structure known as the central helix, which interacts with the other two S2 subunits central helix forming a six-helix bundle (6-HB). This 6-HB allows for

the fusion of the two membranes and the release of the viral genome into the host cell cytoplasm [11]. A schematic representation of the process is shown on Figure 7.

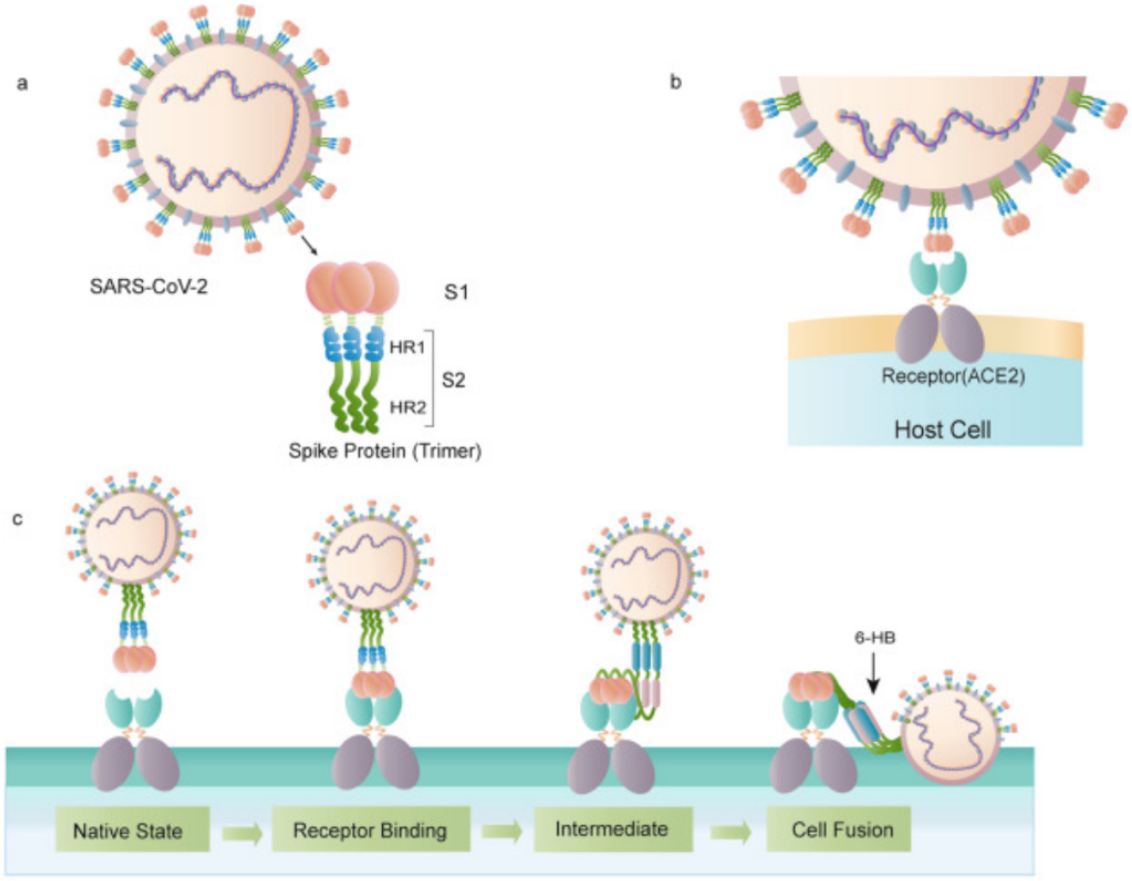

Figure 8: Schematic process of SARS-CoV-2 Spike protein and host cell binding process. A) The Spike protein structure. B) The Spike protein binding to ACE2 receptor. C) The binding and virus-cell fusion process mediated by Spike protein. Adapted from [11]

Once the SARS-CoV-2 has entered the cell, the SARS-CoV-2 (+) gRNA initiates the production of viral proteins forming double-membrane vesicles from endoplasmic reticulum membranes [13]. These vesicles shield the double-stranded RNA that is being generated inside from detection by cytoplasmic pattern recognition receptors (PRRs) that can trigger immunogenic response against it. It is believed that PRRs such as RIG-I and MDA5 are in charge of recognizing long dsRNA and initiate a signalling cascade to promote the transcription of Interferons (IFN) type I and II that will induce antiviral cellular state both autocrine and paracrine [13,14]. In addition, there is cytokine production that promotes the adaptative B cell and T cell responses from interferon stimulated genes that are being triggered because of IFN receptors [13] The mechanism can be seen on Figure 9.

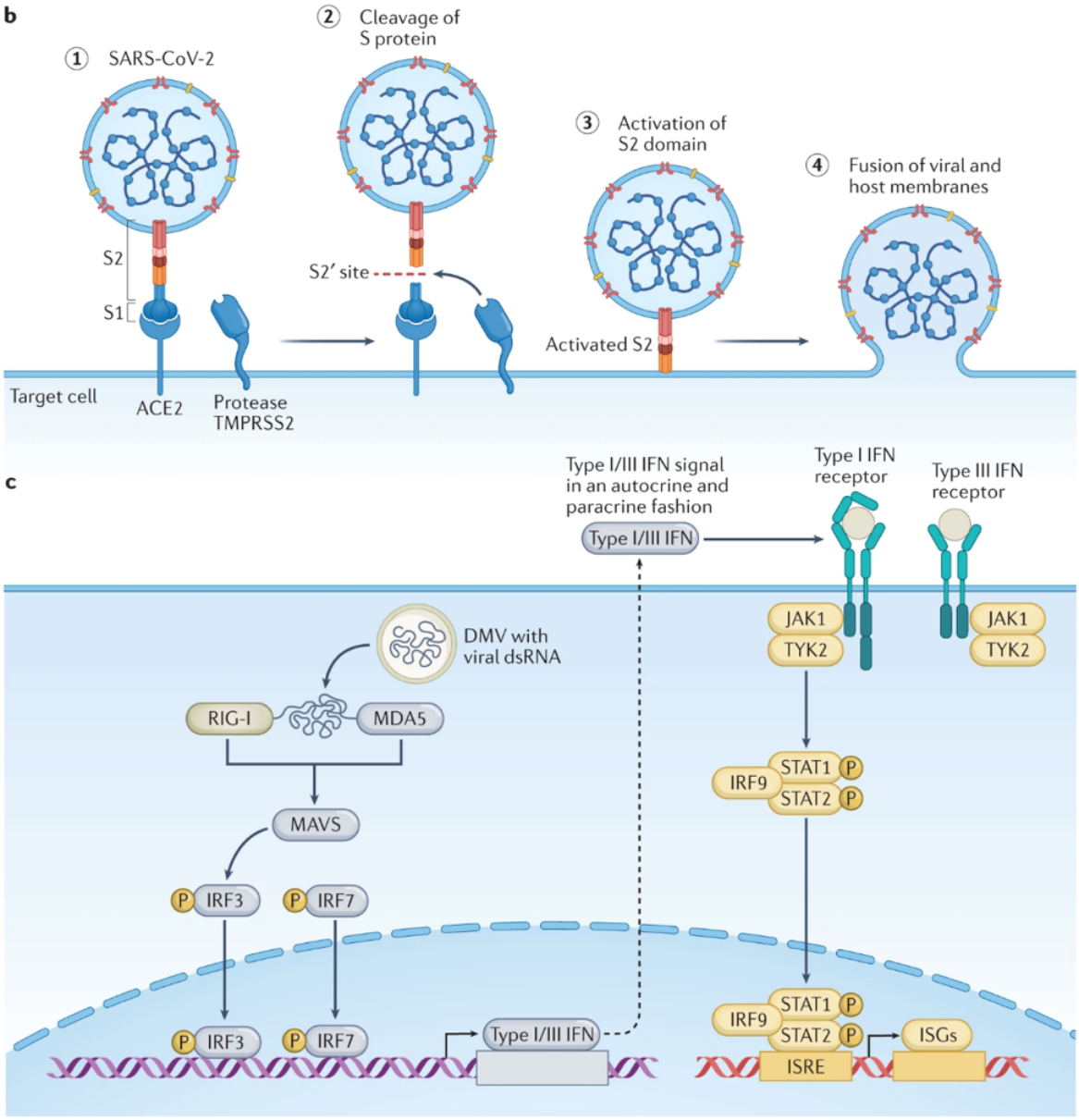

Figure 9: Schematic representation of: B) SARS-CoV-2 and host cell interaction process. C) Host cell viral state to defend against the SARS-CoV-2 infection. Adapted from [13]

The SARS-CoV-2 RNA secondary structure has been systematically characterized by nuclear magnetic resonance and complemented these data with dymethilsulfate footprinting analysis used to identify dynamic regions or regions that can exist in multiple conformations [3,15]. There were identified 15 RNA elements such as stem-loop domains present at the genomic 5'-end and 3'-UTR in addition to RNA constructs corresponding to cis-acting elements from the ORF regions (Figure 10).

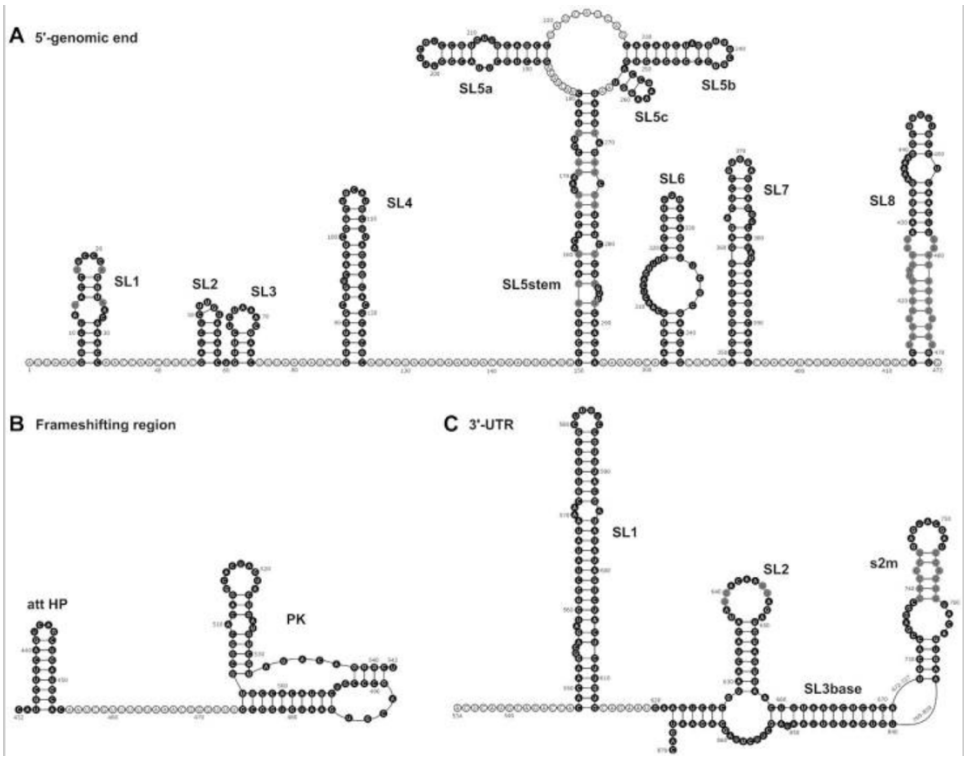

Figure 10: Experimentally derived structures for SARS-CoV-2 genome. A) the cis-elements of the 5'-genomic end. B) the frameshifting region. C) the cis-elements of the 3'UTR. Cis-elements analysed by NMR spectroscopy are highlighted in black and regions with unclear base pairing patterns or high reactivity with DMS are shown in grey. Extracted from [15]

Genomic variability in SARS-CoV-2 variants is produced by mutations and recombination. Most SARS-CoV-2 mutations are expected to be either neutral or mildly deleterious because the fact that a deleterious mutation will potentially prevent the virus from developing. Mutations can also improve virulence, infectivity, or transmissibility due to selective pressure caused by vaccines and antiviral drugs [16,17]. Mutations are not always found because of their beneficial effects but also because of founder effect, which occurs when a mutation is passed on to all the descendants early in the evolution process. Despite the effects on viral fitness, mutations obtained early in the evolution process may be passed on to future descendants because they appear simultaneously with other mutations that confer an advantage to the virus. The SARS-CoV-2 has undergone genetic diversification resulting in the creation of new lineages and variants. The Variants of concern (VOC) are the variants that have indications of higher transmissibility or virulence, a negative impact on epidemiology or a decline in the efficacy of current vaccines or therapies [16,17]. Some of the most known VOCs are Alpha, Beta, Delta, Gamma and Omicron.

The first one to appear was the Alpha VOC in UK in September 2020 and has been detected in many other countries. The Alpha variant has some mutations in the RBD such as N501Y, P681H and 69/70 deletion [18,19]. The Beta VOC has some RBD mutations as well, such as N501Y, E484K and K417N. Beta variant was identified in South Africa in October 2020 and it manifested a very high local prevalence [18]. Gamma VOC was discovered in Brazil in December 2020 and initially it included several mutations including 10 amino acid changes in spike protein, three deletions, four synonymous mutations, and one insertion [18,20]. The Delta VOC was first detected in India early 2021 and contains some several mutations such as L452R. P681R, T19R, R158G, T478K and D950N located in places like the RBD and the Spike protein allowing it to expand rapidly and became the dominant variant worldwide by late 2021 (Figure11) [16,18]. The Omicron VOC was first identified in South Africa in November 2021 and due to the increased transmissibility, it displaced most of the other variants becoming the predominant VOC (Figure 11) [19]. The Omicron variant has been divided into six subvariants BA.1, BA.2, BA.3, BA.4, BA.5 and BA.2.12.1 because they are genetically and antigenically different from each other. Some of the Omicron variant mutations located at the RBD are G339D, D373P, S375F, K417N and N440K [18].

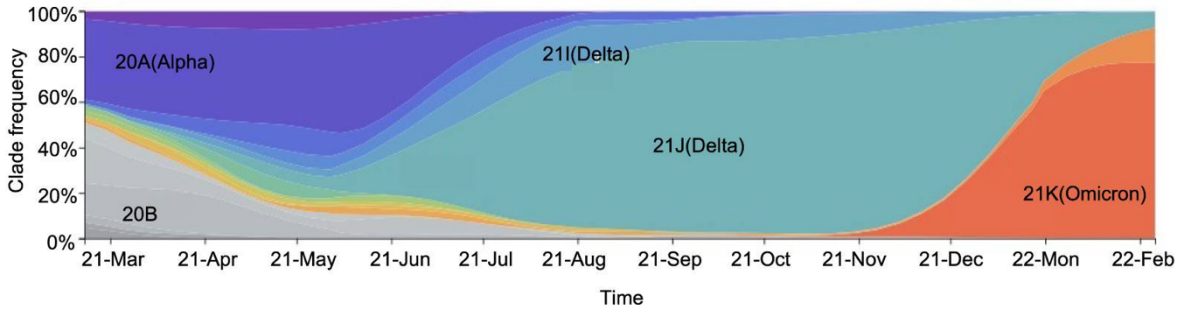

Figure 11: sequential frequency of SARS-CoV-2 VOCs from April 2021 to February 2022. Adapted from [19]

Simultaneous mutations in a genome refer to the occurrence of multiple mutations in different regions of a genome at the same time. In SARS-CoV-2 context, this refers to the emergence of clusters of mutations that can be important in understanding its evolution process. Simultaneous mutations can be caused by a variety of factors, including the maintenance of the protein or RNA structures, selective pressure from the host immune system or drug treatments, as well as founder effects. The biological reasons for the appearance of simultaneous mutations are not fully understood,

hence, it is an interesting topic to research on. The study of clusters of multiple mutations is not new in science. This approach has been used in studies for different diseases such as cancer, checking for example different Single Nucleotide Polymorphisms (SNPs) that may occur simultaneously and affect the probability of developing a cancer or even to respond to different treatments [21]. There are not so many studies in regards of simultaneous mutations or clusters for SARS-CoV-2 but some scientists back in 2020 already established a similar relationship between the virus mutations and its ability to spread. Yang et al [22] studied the evolution route and the transmission of the virus through phylodynamic analysis arriving at the conclusion that there were 4 genetic clusters responsible for the major outbreaks in various countries. Note that the clusters reported had a high mutation rate without coM-promising infectivity [22]. Another study that relates mutation clusters and the SARS-CoV-2 revealed that existed a cluster in Spike protein that could have changed the sensitivity to the BNT162b2 vaccine but in the end, they concluded that this cluster did not affect to the efficacy of the vaccine [23]. Bioinformatics and machine learning have been useful to predict resistance to drugs of different mutations on HIV based on known resistance mutations to various antiretroviral therapies [24]. This approach could also be useful to predict future efficacy of drug treatments or vaccines in SARS-CoV-2 if studied well enough.

OBJECTIVE

Simultaneous mutations is a topic understudied in the virology field and they can provide crucial information about the virus' evolution and its susceptibility to therapeutics and vaccines. The understanding of simultaneous mutations may help control next variants of the SARS-CoV-2 and even other future pandemics that could occur.

The main objective of this project is **to analyse simultaneous mutations in M-pro and Spike proteins of SARS-CoV-2**.

To achieve this objective, there are different tasks to complete such as:

- To process data from SARS-CoV-2 complete genomes and extract the single mutations and simultaneous mutations.
- To automatize the process with Python programming in order to be able to reproduce it with any other protein of SARS-CoV-2 or any other virus.
- To find a biological reason for the occurrence of simultaneous mutations in order to understand their origin.

HYPOTHESIS

There are simultaneous mutations in M-pro and Spike proteins of SARS-CoV-2.

Regarding the biological reason of simultaneous mutations there are three hypotheses to check:

- Simultaneous mutations are related to the structure and function of the M-pro and Spike SARS-CoV-2 proteins.
- Simultaneous mutations are related the founder effect, *i.e.,* mutations being originated early in the evolution process and transmitted together.
- Simultaneous mutations are related to the SARS-CoV-2 secondary RNA structure.

MATERIALS AND METHODS

Single mutation finding

A total of 5,340,569 SARS-CoV-2 sequences downloaded from the GISAID database [25] on 27 June 2022 were analysed. The chosen sequences were obtained from human samples and were ”high coverage” sequences *i.e.,* they meet the following requirements: (1) less than 1% of unidentified bases, (2) less than 0,05% of unique amino acid mutations to eliminate potential sequencing errors, (3) no insertions or deletions, unless verified by the submitter. For each sequence we extracted single mutations, insertions, and deletions, all of them relative to the reference genome (the SARS-CoV-2 sequence NC_045512.2, isolated from Wuhan and submitted to the GenBank database on 17 January 2020). The data extracted from these sequences were Genome identification, Date, VOC, Number of mutations and mutation types for each protein, and the mutation at nucleotide level. From this information we calculated the frequency of M-pro and Spike protein mutations. All the procedure done has been automatized with Python programming to be able to get all single mutations from any SARS-CoV-2 protein from a dataset with the same characteristics.

These proteins were chosen because their properties; the first one is a little protein with not many mutations; the second one is one of the proteins with the highest number of mutations. Studying these two proteins can help us understand simultaneous mutations in any protein indistinctively of their mutation rate.

Simultaneous mutation finding

From the frequency of single non-synonymous mutations, we obtained the mutations that appear in pairs. If two mutations appear simultaneously in more than one sample, it was counted as a mutation pair. Table S1 and Table S4 contain an example of the mutations found in pairs of M-pro and Spike proteins respectively.

To select the mutations that may occur simultaneously, from now on described as simultaneous mutations, we used the rate of co-mutation index (RCM) [26]. This statistical index identifies the co-occurring mutations with equal or close mutation frequencies. The ideal simultaneous mutations would be those that appear almost exclusively in pairs.

RCM is calculated as: $RCM_{A,B} = \frac{M_A \cap M_B}{\sqrt{(M_A \cdot M_B)}}$ where M_x is the number of genomes with the mutation x. RCM ranges from 0 to 1 and when two mutations co-occur, it is 1 or close. We used a relaxed RCM of 0.90 instead of 1 to exclude the effects of sequencing errors. All the procedure done has been automatized with Python programming to be able to get all simultaneous mutations from any SARS-CoV-2 protein from a dataset with the same characteristics.

Simultaneous mutations related to protein structure

We have used the "PDBParser" and "NeighborSearch" modules from the Bio.PDB library to read PDB files containing the protein structure and calculate the distance between a pair of amino acids, using a cut off distance of 100 Å, with Python programming. We have done this to select the mutation pairs that are close enough to make an interaction. The protein structure has been uploaded from the PDB database.

Simultaneous mutations related to RNA structure

We have used a prediction of the SARS-CoV-2 RNA structure obtained with the Vienna RNAFold program [27]. It is based on carefully measured thermodynamics parameters. Algorithms are used to compute ground states, base pairing probabilities, as well as thermodynamic properties in order to determine the most probable interaction between nucleotides [27]. The mutation nucleotides extracted from the data have been compared to the prediction obtained.

RESULTS AND DISCUSSION

Simultaneous mutation finding

From 922,474 non-synonymous M-pro mutations, we found a total of 486 pairs of mutations that appear at least 10 times together in the M-pro protein. After screening the results by the RCM index, only 4 mutation pairs had an RCM index over 0.90 (see Table 1 and Table S1).

Table 1: M-pro pair mutations with a RCM>0.90

Mutation pairs	Nr* Mut1 & Mut2	Mut1	Nr* mut1	Mut2	Nr* mut2	RCM
T225M T226S	415	T225M	415	T226S	416	0.998797
G29C R105L	57	G29C	59	R105L	59	0.966102
R105L A210S	57	R105L	59	A210S	64	0.927596
G29C A210S	57	G29C	59	A210S	64	0.927596

*Nr stands for the number of times each mutation is observed (mut 1 or mut 2) or both mutations simultaneously (mut1 & mut2).

From 44,278,219 non-synonymous Spike mutations we have done the same procedure, getting 253 mutation pairs that occur at least 1000 times together and have higher RCM than 0.90 (see Table S4).

These indicates that there are mutations that do appear almost exclusively alongside others in M-pro and Spike proteins. So, **we can conclude that simultaneous mutations do exist in the SARS-CoV-2 M-pro and Spike proteins.**

Simultaneous mutations related to protein structure

To test if mutation pairs are related to the protein structure, we have checked the distance in the protein 3D-structure between the two amino acids of the mutation pair. If the distance discovered in the M-pro and Spike proteins mutation pairs is 5 Å approximately or less apart, these amino acids could interact. However, the distance between the pairs in M-pro were all greater than 17 Å, with the exemption of the T225M and T226S pair which are contiguous amino acids. Having done the same process with the mutation pairs in Spike the results found were anything from different. There were only 8 mutation pairs with less than 5 Å of distance between them (Table 2 and Table S4) which 6 of them are contiguous ones.

Table 2: Spike pair mutations with RCM>0.90 and distance between amino acids < 5Å

Mutation pairs	Nr* Mut1 & Mut2	RCM	Distance(Å)
Y144S Y145N	5494	0.92	1.31
S375F T376A	327387	0.96	1.32
R34P G35R	164	0.92	1.33
D405N R408S	322271	0.99	3.48
S371F S373P	324718	0.95	3.92
S373P S375F	356221	0.99	4.07
D405N Y505H	323172	0.95	4.58
T547K L981F	28567	0.97	4.88

*Nr stands for number of times that each simultaneous mutation is observed

The amino acids involved in the simultaneous mutations are too far apart to interact, and we can **conclude that the amino acids implied in simultaneous mutations do not directly interact on the protein structure**. On the other hand, simultaneous mutations can be involved in protein structure maintenance indirectly with other amino acids that are not mutated *i.e*, one mutation causes the loss of a non-covalent interaction with an amino acid not mutated and the second mutation re-establishes a similar interaction with the same or a near amino acid in the protein structure.

Simultaneous mutations related to functionality

Another explanation for simultaneous mutations could be the loss or gain of function. If any mutation pair would be beneficial for the virus, it is to be expected to repeat these mutations through different virus variants and not only spread from just an initial one. Therefore, there could be some of them that cause a loss of function and develop into a weaker form of the virus or on the contrary, it could cause a gain of function and develop into stronger versions of the virus. To check this, different articles have been looked up to in regards of mutations that can increase or decrease function of M-pro and Spike proteins.

Julia M Flynn, et al. [28] performed a comprehensive mutational scan of the protease M-pro that analysed the function of all the possible single amino acid changes. The results can be found at Figure 11. T225M, T226S, R105L are three of the individual mutations found on the M-pro simultaneous mutations. They do not show any significant loss of function on the heatmap. On the contrary, M-pro proteins with G29C

and A210S mutations lose some of their activity. These data do not clarify if simultaneous mutations are related to the M-pro protein function. One explanation for this could be that those viruses that get the impaired mutations fail because its lack of function. Moreover, if they also get the mutations with gain of function, it could compensate the prior loss and consequently develop as usual or with higher rate.

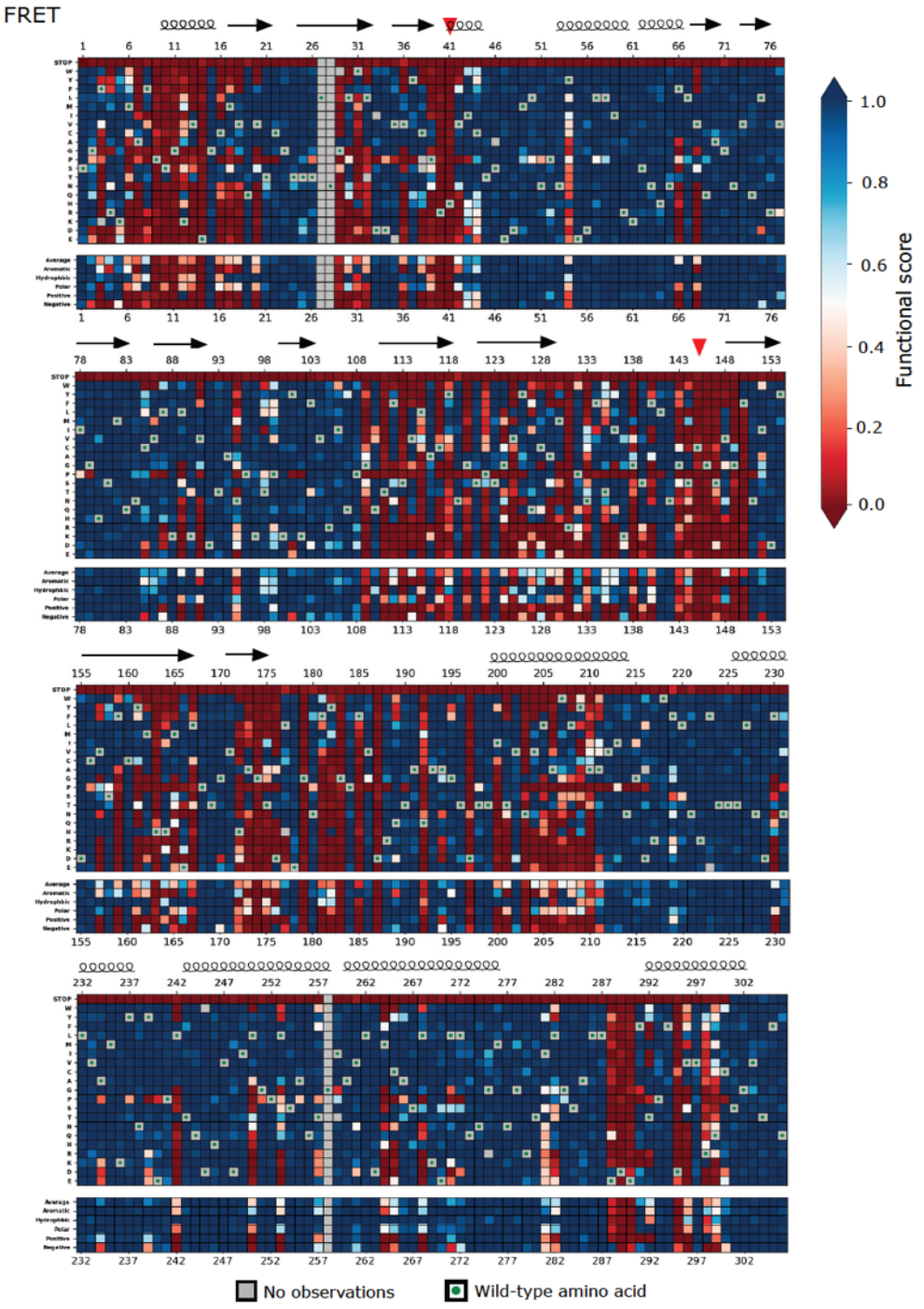

Figure 11: Heatmap representation of the main protease (M-pro) functional scores measured in the fluorescence resonance energy transfer (FRET) screen.

On what Spike concerns, there are multiple mutations found on the mutation pairs finding with an RCM over 0.90. We have grouped the mutations in clusters generated by checking if one mutation is also encountered in another pair with high RCM and checking again if the third new mutation also pairs with the second one. Having done this with all the 253 pairs of Spike protein we found three clusters (Table 3 and Tables S5-7).

Table 3: SARS-CoV-2 Spike protein clusters found

SARS-CoV-2 Spike protein clusters
D796Y, E484A, G339D, N440K, N679K, N969K, Q498R, Q954H, S373P, S375F, S477N, Y505H
D950N, L452R, P681R, T19R, T478K
A570D, D1118H, S982A, T716I

The signature mutations of different SARS-CoV-2 VOCs are shown in Figure 12 and if checked with the clusters found and discussed above, the individual mutations found in the clusters appear to be mostly signature mutations.

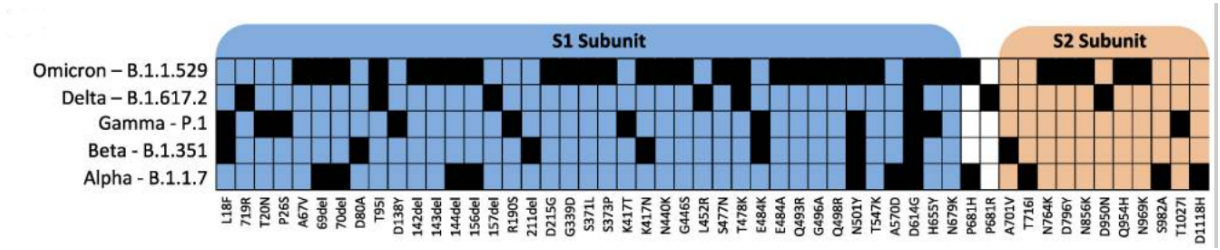

Figure 12: SARS-CoV-2 Spike mutations through different VOCs. Adapted from [38]

Inside the first cluster we can find D796Y, E484A, G339D, N440K, N679K, N969K, Q498R, Q954H, S373P, S375F, S477N and Y505H. Some of the mutations did not appear on the cluster initially because the RCM of the pair was close to 0.90 but not over the threshold. This was solved by checking manually if the mutation pair existed alongside of good metrics regarding the number of times that appeared. Those are signature mutations of the Omicron variant found in the NTD region, in the RBD, around the furin cleavage region, in the FP domain and HR1 regions [29]. These amino acid changes in Omicron are related to an enhanced immunity scape because the mutations in NTD region (a supersite of recognition by antibodies) make the virus more efficient in escaping from the immune system and reinfect people. Besides, these changes also are involved in stabilizing the S1 to allow strong interactions between

RBD and ACE2 converting the virus in a more infectious variant [29]. E484A, Q498R and S477N are mutations located in the RBD involved in the escape of neutralizing antibodies [29]. K417N is a mutation shared among different variants such as Omicron, Delta, and Gamma but as it is paired with Omicron signature variant mutations, the pairs where it appears are mostly Omicron VOCs [29]. N679K also reinforces the higher transmissibility because the substitution of an asparagine for a positively charged lysine residue, making the environment prone for the furin cleavage [29]. D405N, R408S and S371F mutations are found also in Omicron variants and they are also involved in eliminating the neutralizing capacity of antibodies [29]. T19I does not affect known functional domains but it has been seen that in presence of N856K and other mutations such the three serine residues in a small loop S371F, S373P, S375F and T375A changes, produce a severely impaired S-mediated infection [30]. Q493R is not an Omicron signature variant mutation but it is found in a key place for the interaction between RBD and ACE2 [30]. It was found that this change is tolerated and does not affect the linkage. However, it may affect the interaction between the virus and some SARS-CoV-2 Neutralizing Antibodies such as LY-CoV555 and LY-CoV016 making these monoclonal antibodies therapy less efficient [31]. All the mutations exposed above could demonstrate why the Omicron variant and consequent subvariants are two to three times as infectious as the Delta variant [32]. In addition to this, this could explain why the Omicron variant is 15-80% less likely to cause a lethal disease [33,34]. The acquisition of mutations that serve the purpose of escaping the immune system alongside the ability of infect people fully vaccinated make the virus more dominant and fast spreading than other variants [32]. On the other hand, the likely loss of S-mediated infection could make the virus less harmful for the organisms lowering the replication rate or even deactivate some entrance to the cell pathways [35].

Inside the second cluster we can find D950N, L452R, P681R, T19R and T478K. D950N is located at the S2 region and contributes to the regulation of Spike protein dynamics [36]. L452R and T478K are mutations found in the epitope region of the RBD, indicating that their apparition can be related to the immunity scape [36]. P681R is located at the furin-cleavage site and may enhance the activity of the spike protein and the ACE2 binding [36,37]. T19R is a mutation located in the NTD and it is also related to immunity scape because the NTD is a region targeted by most anti-NTD neutralizing antibodies [36]. These mutations are Delta variant signature mutations

[36,38] and, as the Omicron variant signature mutations, they serve the purpose of improving the virus overall. The founder effect could explain why there are some mutations discovered in Delta but are not signature mutations of the variant.

Inside the third cluster we can find A570D, D1118H, S982A and T716I, Alpha variant signature mutations located mainly on S1 and S2 subunits [39]. S982A substitution enhances the ''up'' RBD state by removing the interaction with T547 and A570D can form a hydrogen bond with N856. These two mutations together re-establish the ''down'' RDB state, reinforcing the stacking of the S1 subdomain loop against the HR1 helix [31,33]. Similarly, whereas the T716I mutation disrupts a hydrogen bond, the D1118H appears to play a stabilizing role allowing the Spike trimer to form a histidine triad that in overall stabilizes the protein [31,33]. In this case, the Alpha variant signature mutations are related to maintain the structure of the Spike protein and thus maintain its functionality.

From the results presented above, we **can conclude that simultaneous mutations can be grouped in clusters and the clusters are related to a specific VOC.** The early acquisition of this signature mutations has made them to transmit along together through the next generations and allow us to identify the virus' VOC. This phenomenon is also called the founder effect. Individual mutations found in the clusters can be either beneficial or deleterious for the virus but on the whole, simultaneous mutations on the Spike protein provide the virus with an advantage that allows it to expand easily and generate a new variant.

The study of simultaneous mutations can be beneficial to understand the apparition of next VOCs and advance ourselves to the harmful effects that these new mutations can produce. The understanding of simultaneous mutations can also help to adjust parameters at designing drugs or vaccines that can be useful to cover all the variants possible. Moreover, the re-design of pre-existing drugs or vaccines is easier if the simultaneous mutations that generate the new VOC are identified.

Simultaneous mutations related to RNA structure

Individual mutation's nucleotides have been extracted (Table S3 and Table S8) and have been compared to the prediction made by the Vienna RNAFold program. There is also added the RCM parameter to determine if the mutation pair is found exclusively

in pairs or not. The interaction nucleotide prediction stands for the nucleotide that would bind the nucleotide of the mutation 1. Results are shown on Tables 4 and 5.

Table 4: mutation pairs with a RCM index >0.90 and their prediction nucleotide of M-pro protein

Mutation pairs	RCM	Nucleotide mut1	Nucleotide mut2	Interaction nucleotide prediction
T225M T226S	0.998797	10,728	10,730	10,705
G29C R105L	0.966102	10,139	10,368	9,976
R105L A210S	0.927596	10,368	10,682	*
G29C A210S	0.927596	10,139	10,682	9,976

*Empty spaces are caused by the absence of an interaction in the prediction made by the Vienna RNAFold program. This could happen because the region is not paired or it is a dynamic zone.

Table 5: mutation pairs with >0.90RCM index and closest prediction possible of the Spike protein

Mutation pairs	RCM	Nucleotide mut1	Nucleotide mut2	Interaction nucleotide prediction
Q498R Y505H	0,992651	23,055	23,075	23,071
S371F T376A	0,990298	22,674	22,688	22,704
S371F S375F	0,951463	22,674	22,686	22,704
Q954H N969K	0,998647	24,424	24,469	24,491
S371F S373P	0,951816	22,674	22,679	22,704
Y144S Y145N	0,919833	21,993	21,995	21,970

For example, in M-pro protein G29C R105L simultaneous mutation, first mutation G29C nucleotide is 10,139 should bind to the nucleotide 9,976 but the second mutation R105L nucleotide is 10,368.

Comparing all the nucleotides exposed above, we have seen that there is no relationship between simultaneous mutations and RNA structure. This allows us **to conclude that simultaneous mutations are not related to the secondary RNA structure on SARS-CoV-2.**

CONCLUSIONS

Based on the results obtained from the analysis of 922,474 M-pro protein mutations and 44,278,219 Spike protein mutations of SARS-CoV-2, we can conclude that **simultaneous mutations do exist in these proteins**.

Having checked the distance between the amino acids that form the simultaneous mutations, we conclude that **the amino acids implied in simultaneous mutations do not directly interact on the protein structure**. Nonetheless, there is scientific evidence that some simultaneous mutations are related to maintain certain hydrogen bonds that allow the protein to still have a functional conformation. The atoms involved in this new linkage are not specifically the two involved in the simultaneous mutations. In addition, we have been able to assess that **simultaneous mutations are not related to the secondary RNA structure** using a prediction made by the Vienna RNAFold program.

The biological function is altered by the individual mutations found on M-pro and Spike proteins. In both proteins, these mutations can lead to an increase or decrease of the protein function but overall, they provide the virus an advantage and allow it to develop in a more optimal way. The grouping of the simultaneous mutations in clusters have allowed the scientist to classify the new emerging versions of SARS-CoV-2 in Variants of Concern (VOCs). Early detection of new VOCs is crucial to controlling the spread of infectious diseases, especially those caused by high mutation rates viruses like SARS-CoV-2. After generating three different Spike protein clusters related to three SARS-CoV-2 VOCs, we are able to conclude that in SARS-CoV-2 Spike protein, **simultaneous mutations can be grouped in clusters and the clusters are related to the founder effect,** a phenomenon that occurs when some mutations happen together, and they are transmitted along because the beneficial effects on the whole. Some of the advantages of studying simultaneous mutations could be the prevention of the spread of new VOCs implementing targeted testing or re-designing vaccines and treatments to limit the transmission of the virus. This can lessen the negative effects of newly developing infectious diseases on human health and prevent future pandemics.

All of this procedure has successfully been automated with Python programming in order to check the same parameters with other SARS-CoV-2 proteins or any other virus proteins whose data is presented in the same format.

BIBLIOGRAPHY

[1] Dhama K, Khan S, Tiwari R, Sircar S, Bhat S, Malik YS, et al. Coronavirus Disease 2019-COVID-19. Clin Microbiol Rev 2020;33:e00028-20. https://doi.org/10.1128/CMR.00028-20.

[2] Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 2019;17:181–92. https://doi.org/10.1038/s41579-018-0118-9.

[3] Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. Cells 2020;9:1267. https://doi.org/10.3390/cells9051267.

[4] Eriani G, Martin F. Viral and cellular translation during SARS-CoV-2 infection. FEBS Open Bio 2022;12:1584–601. https://doi.org/10.1002/2211-5463.13413.

[5] Tabibzadeh A, Esghaei M, Soltani S, Yousefi P, Taherizadeh M, Safarnezhad Tameshkel F, et al. Evolutionary study of COVID-19, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as an emerging coronavirus: Phylogenetic analysis and literature review. Vet Med Sci 2021;7:559–71. https://doi.org/10.1002/vms3.394.

[6] Long S. SARS-CoV-2 Subgenomic RNAs: Characterization, Utility, and Perspectives. Viruses 2021;13:1923. https://doi.org/10.3390/v13101923.

[7] Gorkhali R, Koirala P, Rijal S, Mainali A, Baral A, Bhattarai HK. Structure and Function of Major SARS-CoV-2 and SARS-CoV Proteins. Bioinform Biol Insights 2021;15:117793222110258. https://doi.org/10.1177/11779322211025876.

[8] Hu Q, Xiong Y, Zhu G-H, Zhang Y-N, Zhang Y-W, Huang P, et al. The SARS-CoV-2 main protease (Mpro): Structure, function, and emerging therapies for COVID-19. MedComm (Beijing) 2022;3:e151. https://doi.org/10.1002/mco2.151.

[9] Kneller DW, Phillips G, O'Neill HM, Jedrzejczak R, Stols L, Langan P, et al. Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography. Nat Commun 2020;11:3202. https://doi.org/10.1038/s41467-020-16954-7.

[10] Saldivar-Espinoza B, Macip G, Pujadas G, Garcia-Vallve S. Could nucleocapsid be a next-generation COVID-19 vaccine candidate? International Journal of Infectious Diseases 2022;125:231–2. https://doi.org/10.1016/j.ijid.2022.11.002.

[11] Huang Y, Yang C, Xu X, Xu W, Liu S. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin 2020;41:1141–9. https://doi.org/10.1038/s41401-020-0485-4.

[12] Jackson CB, Farzan M, Chen B, Choe H. Mechanisms of SARS-CoV-2 entry into cells. Nat Rev Mol Cell Biol 2022;23:3–20. https://doi.org/10.1038/s41580-021-00418-x.

[13] Lamers MM, Haagmans BL. SARS-CoV-2 pathogenesis. Nat Rev Microbiol 2022;20:270–84. https://doi.org/10.1038/s41579-022-00713-0.

[14] Yin X, Riva L, Pu Y, Martin-Sancho L, Kanamune J, Yamamoto Y, et al. MDA5 Governs the Innate Immune Response to SARS-CoV-2 in Lung Epithelial Cells. Cell Rep 2021;34:108628. https://doi.org/10.1016/j.celrep.2020.108628.

[15] Wacker A, Weigand JE, Akabayov SR, Altincekic N, Bains JK, Banijamali E, et al. Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. Nucleic Acids Res 2020;48:12415–35. https://doi.org/10.1093/nar/gkaa1013.

[16] Saldivar-Espinoza B, Garcia-Segura P, Novau-Ferré N, Macip G, Martínez R, Puigbò P, et al. The Mutational Landscape of SARS-CoV-2. Int J Mol Sci 2023;24:9072. https://doi.org/10.3390/ijms24109072.

[17] Choi JY, Smith DM. SARS-CoV-2 Variants of Concern. Yonsei Med J 2021;62:961. https://doi.org/10.3349/ymj.2021.62.11.961.

[18] Chen K-WK, Tsung-Ning Huang D, Huang L-M. SARS-CoV-2 variants - Evolution, spike protein, and vaccines. Biomed J 2022;45:573–9. https://doi.org/10.1016/j.bj.2022.04.006.

[19] Sun C, Xie C, Bu G-L, Zhong L-Y, Zeng M-S. Molecular characteristics, immune evasion, and impact of SARS-CoV-2 variants. Signal Transduct Target Ther 2022;7:202. https://doi.org/10.1038/s41392-022-01039-2.

[20] Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science (1979) 2021;372:815–21. https://doi.org/10.1126/science.abh2644.

[21] Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, de Borja R, et al. Comprehensive Analysis of Hypermutation in Human Cancer. Cell 2017;171:1042-1056.e10. https://doi.org/10.1016/j.cell.2017.09.048.

[22] Yang X, Dong N, Chan EW-C, Chen S. Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries. Emerg Microbes Infect 2020;9:1287–99. https://doi.org/10.1080/22221751.2020.1773745.

[23] Messali S, Bertelli A, Campisi G, Zani A, Ciccozzi M, Caruso A, et al. A cluster of the new SARS-CoV-2 B.1.621 lineage in Italy and sensitivity of the viral isolate to the BNT162b2 vaccine. J Med Virol 2021;93:6468–70. https://doi.org/10.1002/jmv.27247.

[24] Blassel L, Zhukova A, Villabona-Arenas CJ, Atkins KE, Hué S, Gascuel O. Drug resistance mutations in HIV: new bioinformatics approaches and challenges. Curr Opin Virol 2021;51:56–64. https://doi.org/10.1016/j.coviro.2021.09.009.

[25] GISAID - gisaid.org n.d. https://gisaid.org/ (accessed May 1, 2023).

[26] Qin L, Ding X, Li Y, Chen Q, Meng J, Jiang T. Co-mutation modules capture the evolution and transmission patterns of SARS-CoV-2. Brief Bioinform 2021;22. https://doi.org/10.1093/BIB/BBAB222.

[27] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms for Molecular Biology 2011;6:26. https://doi.org/10.1186/1748-7188-6-26.

[28] Flynn JM, Samant N, Schneider-Nachum G, Barkan DT, Yilmaz NK, Schiffer CA, et al. Comprehensive fitness landscape of SARS-CoV-2 Mpro reveals insights into viral resistance mechanisms. Elife 2022;11:e77433. https://doi.org/10.7554/eLife.77433.

[29] Souza PFN, Mesquita FP, Amaral JL, Landim PGC, Lima KRP, Costa MB, et al. The spike glycoprotein of SARS-CoV-2: A review of how mutations of spike glycoproteins have driven the emergence of variants with high transmissibility and immune escape. Int J Biol Macromol 2022;208:105–25. https://doi.org/10.1016/j.ijbiomac.2022.03.058.

[30] Pastorio C, Zech F, Noettger S, Jung C, Jacob T, Sanderson T, et al. Determinants of Spike infectivity, processing, and neutralization in SARS-CoV-2 Omicron subvariants BA.1 and BA.2. Cell Host Microbe 2022;30:1255-1268.e5. https://doi.org/10.1016/j.chom.2022.07.006.

[31] Starr TN, Greaney AJ, Dingens AS, Bloom JD. Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail

with LY-CoV016. Cell Rep Med 2021;2:100255. https://doi.org/10.1016/j.xcrm.2021.100255.

[32] Chen J, Wang R, Gilby NB, Wei G-W. Omicron Variant (B.1.1.529): Infectivity, Vaccine Breakthrough, and Antibody Resistance. J Chem Inf Model 2022;62:412–22. https://doi.org/10.1021/acs.jcim.1c01451.

[33] Wolter N, Jassat W, Walaza S, Welch R, Moultrie H, Groome M, et al. Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: a data linkage study. Lancet 2022;399:437–46. https://doi.org/10.1016/S0140-6736(22)00017-4.

[34] Christie B. Covid-19: Early studies give hope omicron is milder than other variants. BMJ 2021;375:n3144. https://doi.org/10.1136/bmj.n3144.

[35] Shuai H, Chan JF-W, Hu B, Chai Y, Yuen TT-T, Yin F, et al. Attenuated replication and pathogenicity of SARS-CoV-2 B.1.1.529 Omicron. Nature 2022;603:693–9. https://doi.org/10.1038/s41586-022-04442-5.

[36] Planas D, Veyer D, Baidaliuk A, Staropoli I, Guivel-Benhassine F, Rajah MM, et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. Nature 2021;596:276–80. https://doi.org/10.1038/s41586-021-03777-9.

[37] Cherian S, Potdar V, Jadhav S, Yadav P, Gupta N, Das M, et al. SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. Microorganisms 2021;9:1542. https://doi.org/10.3390/microorganisms9071542.

[38] Magazine N, Zhang T, Wu Y, McGee MC, Veggiani G, Huang W. Mutations and Evolution of the SARS-CoV-2 Spike Protein. Viruses 2022;14:640. https://doi.org/10.3390/v14030640.

[39] Scovino AM, Dahab EC, Vieira GF, Freire-de-Lima L, Freire-de-Lima CG, Morrot A. SARS-CoV-2's Variants of Concern: A Brief Characterization. Front Immunol 2022;13:834098. https://doi.org/10.3389/fimmu.2022.834098.

SUPLEMENTARY DATA

Table S1: Example of M-pro protein data extracted from the 5,340,569 SARS-CoV-2 sequences downloaded from the GISAID database. Simultaneous mutations and their rate of co-mutation (RCM) and the distance between mutation 1 and mutation 2.

Simultaneous mutations	Nr*	Mut1	Nr* Mut1	Mut2	Nr* Mut2	RCM	Distance between Mut1 and Mut2 (Å)
T225M T226S	415	T225M	415	T226S	416	0.998797	1.33
G29C R105L	57	G29C	59	R105L	59	0.966102	17.61
R105L A210S	57	R105L	59	A210S	64	0.927596	24.13
G29C A210S	57	G29C	59	A210S	64	0.927596	33.78
S46A S158T	148	S46A	210	S158T	183	0.754964	30.89

*Nr stands for number of times that each simultaneous mutation or individual mutation is observed

Table S2: Example of M-pro protein data extracted from the 5,340,569 SARS-CoV-2 sequences downloaded from the GISAID database. Simultaneous mutations and the VOC where they appear.

Simultaneous mutations	VOC
T225M T226S	Alpha: 415
G29C R105L	Omicron: 26, Delta: 29, Other: 2
R105L A210S	Omicron: 26, Delta: 29, Other: 2
G29C A210S	Omicron: 26, Delta: 29, Other: 2
S46A S158T	Delta: 148

Table S3: Example of M-pro protein data extracted from the 5,340,569 SARS-CoV-2 sequences downloaded from the GISAID database. Simultaneous mutations and their respective mutations at nucleotide level.

Simultaneous mutations	Nucleotide Mut1	Nucleotide Mut2
T225M T226S	C10728T_A10729G	A10730T
G29C R105L	G10139T	G10368T
R105L A210S	G10368T	G10682T
G29C A210S	G10139T	G10682T
S46A S158T	T10190G	T10526A

Table S4: Example of Spike protein data extracted from the 5,340,569 SARS-CoV-2 sequences downloaded from the GISAID database. Simultaneous mutations and their rate of co-mutation (RCM) and the distance between mutation 1 and mutation 2.

Simultaneous mutations	Nr*	Mut1	Nr* Mut1	Mut2	Nr* Mut2	RCM	Distance between Mut1 and Mut2 (Å)
Q954H N969K	359,689	Q954H	360,297	N969K	360,056	0.998647	21.97
S373P S375F	356,221	S373P	357,802	S375F	356,651	0.997187	4.07
A570D S982A	914,077	A570D	916,865	S982A	916,572	0997119	15.44
N764K Q954H	358,217	N764K	358,731	Q954H	360,297	0.996395	16.63
N764K N969K	358,054	N764K	358,731	N969K	360,056	0.996275	21.94
A570D D1118H	912,607	A570D	916,865	D1118H	915,237	0.996241	79.05
S982A D1118H	912,376	S982A	916,572	D1118H	915,237	0.996148	95.98
V213G D405N	327,971	V213G	329,384	D405N	329,446	0.995616	66.93
T376A D405N	326,955	T376A	327,589	D405N	329,446	0.995248	9.39

*Nr stands for number of times that each simultaneous mutation or individual mutation is observed

Table S5: Example of Spike protein data extracted from the 5,340,569 SARS-CoV-2 sequences downloaded from the GISAID database. Simultaneous mutations and the VOC where they appear. The example contains de Delta variant signature mutations.

Simultaneous mutations	VOC
T19R L452R	Delta: 3,061,750, Other: 328, Alpha: 17, Omicron: 70
T19R T478K	Delta: 3,065,199, Other: 267, Omicron: 225, Alpha: 15
T19R P681R	Delta: 3,068,108, Other: 503, Omicron: 66, Alpha: 7
T19R D950N	Delta: 2,952,404, Other: 1284, Omicron: 51, Alpha: 6
L452R T478K	Delta: 3,072,601, Alpha: 59, Other: 1,211, Omicron: 1,820, Beta: 16
L452R P681R	Delta: 3,070,502, Other: 6640, Omicron: 111, Alpha: 9, Beta: 2
L452R D950N	Delta: 2,954,555, Alpha: 12, Other: 296, Omicron: 51, Beta: 2
T478K P681R	Delta: 3,073,800, Other: 504, Omicron: 367, Alpha: 11, Beta: 2
T478K D950N	Delta: 2,958,124, Alpha: 12, Other: 289, Omicron: 159, Beta: 2
P681R D950N	Delta: 2,962,516, Other: 493, Omicron: 55, Alpha: 11, Beta: 2

Table S6: Example of Spike protein data extracted from the 5,340,569 SARS-CoV-2 sequences downloaded from the GISAID database. Simultaneous mutations and the VOC where they appear. The example contains Alpha variant signature mutations.

Simultaneous mutations	VOC
A570D T716I	Alpha: 913,406, Other: 185, Delta: 5, Beta: 2
A570D S982A	Alpha: 913,726, Other: 325, Delta: 23, Beta: 3
A570D D1118H	Alpha: 912,439, Other: 154, Delta: 11, Beta: 3
T716I S982A	Alpha: 913,601, Other: 473, Beta: 5, Delta: 13
T716I D1118H	Alpha: 911,976, Other: 366, Delta: 19, Beta: 8
S982A D1118H	Alpha: 912,161, Other: 206, Delta: 6, Beta: 3

Table S7: Example of Spike protein data extracted from the 5,340,569 SARS-CoV-2 sequences downloaded from the GISAID database. Simultaneous mutations and the VOC where they appear. The example contains some of the Omicron variant signature mutations.

Simultaneous mutations	VOC
T19I V213G	Omicron: 325,633, Other: 87
T19I G339D	Omicron: 323,916, Other: 91
T19I S371F	Omicron: 320,965, Other: 120
T19I S373P	Omicron: 324,966, Other: 125
T19I S375F	Omicron: 324,355, Other: 119
T19I T376A	Omicron: 323,920, Other: 118
T19I D405N	Omicron: 325,566, Other: 89
T19I R408S	Omicron: 318,996, Other: 85
T19I K417N	Omicron: 323,421, 'Beta': 178, Other: 92, Delta: 1
T19I N440K	Omicron: 306,184, Other: 101
T19I E484A	Omicron: 323,102, Delta: 16, Other: 78
T19I Q493R	Omicron: 321,939, Other: 74
T19I Q498R	Omicron: 319,271, Other: 71
T19I Y505H	Omicron: 320,509, Other: 78
T19I N679K	Omicron: 325,570, Other: 120, Delta: 2
T19I N764K	Omicron: 326,132, Other: 96
T19I D796Y	Omicron: 326,100, Other: 95, Delta: 5
T19I Q954H	Omicron: 326,574, Other: 124, Delta: 1
T19I N969K	Omicron: 326,405, Other: 108, Delta: 1
V213G G339D	Omicron: 326,571, Other: 100, Delta: 1
V213G S371F	Omicron: 323,256, Other: 83, Delta: 1
V213G S373P	Omicron: 327,400, Other: 86, Delta: 1
V213G S375F	Omicron: 327,016, Other: 85, Delta: 1
V213G T376A	Omicron: 326,366, Other: 75, Delta: 1
V213G D405N	Omicron: 327,885, Other: 85, Delta: 1

Table S8: Example of Spike protein data extracted from the 5,340,569 SARS-CoV-2 sequences downloaded from the GISAID database. Simultaneous mutations and their respective mutations at nucleotide level.

Simultaneous mutations	Nucleotide Mut1	Nucleotide Mut2
Q954H N969K	A24424C, A24424T	T24469A, T24469G
S373P S375F	T22679C_A22681G, T22679C_A22681C, T22679C_A22681T, T22679C	C22686T, C22686T_C22687T
A570D S982A	C23271A_T23272C, C23271A	T24506G_A24508G, T24506G
N764K Q954H	C23854G, C23854A	A24424C, A24424T
N764K N969K	C23854G, C23854A	T24469A, T24469G
A570D D1118H	C23271A_T23272C, C23271A	G24914C, G24914C_C24916T
S982A D1118H	T24506G_A24508G, T24506G	G24914C, G24914C_C24916T
V213G D405N	T22200G_G22201T, T22200G, T22200G_G22201A	G22775A, G22775A_T22777C
T376A D405N	A22688G, A22688G_T22690A, A22688G_T22690C	G22775A, G22775A_T22777C