UNIVERSITAT ROVIRA i VIRGILI

SARS-CoV-2 mutational profiling: an insight into the biological importance of mutation hotspots and coldspots of the main protease (M-pro)

Pol Garcia Segura

TREBALL FINAL DE GRAU BIOTECNOLOGIA

Tutor acadèmic: Dr. Gerard Pujadas Anguiano, Departament de Bioquímica i Biotecnologia, URV (gerard.pujadas@urv.cat)

En cooperació amb: Grup de recerca en Quimioinformàtica i Nutrició (QiN), Departament de Bioquímica i Biotecnologia, URV.

Supervisors: Dr. Santi Garcia-Vallvé (santi.garcia-vallve@urv.cat) i Dr. Gerard Pujadas Anguiano (gerard.pujadas@urv.cat), Departament de Bioquímica i Biotecnologia, URV.

Juny 2021

The truth, however ugly in itself,
is always curious and beautiful to the seeker after it.
Hercule Poirot

Jo, Pol Garcia Segura, amb DNI 47699996-C, sóc coneixedor de la guia de prevenció del plagi a la URV Prevenció, detecció i tractament del plagi en la docència: guia per a estudiants (aprovada el juliol 2017) (http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/) i afirmo que aquest TFG no constitueix cap de les conductes considerades com a plagi per la URV.

Tarragona, 7 de juny de 2021

POL GARCIA SEGURA - DNI 47699996C

Firmado digitalmente por POL GARCIA SEGURA - DNI 47699996C
Fecha: 2021.06.07 09:12:20 +02'00'

(signatura)

TABLE OF CONTENTS

The present work has been developed in the Cheminformatics and Nutrition (QiN) research group of the Biochemistry and Biotechnology Department at Rovira i Virgili University (URV). The focus of the research of the QiN group is the use of computational tools to the development of inhibitors and/or repurposing of existing molecules or drugs to specific targets. Due to the COVID-19 outbreak, the QiN research group has started a new research line with the SARS-CoV-2 main protease as the main target to be eventually inhibited. This work is part of this research line and is directed to gain insight of important structural features of the SARS-CoV-2 main protease.

ABSTRACT

SARS-CoV-2 and the COVID-19 pandemic have marked a milestone in the history of scientific research worldwide. Despite the enormous work to fight against this pandemic and the recent development of effective vaccines, no cure is yet available. Nonetheless, thousands of investigations are being conducted in the entire globe to find a treatment. To ensure the success of such treatments in a mid-long term, it is crucial to characterize SARS-CoV-2 mutations as they might lead to viral resistance. SARS-CoV-2 depends on its main protease (M-pro) to achieve a successful replication within the host cell. Here, a mutation profiling of more than 200,000 genomes of SARS-CoV-2 is made. Also, due to its pivotal role in viral resistance, a more detailed analysis of SARS-CoV-2 M-pro is conducted in order to widen our knowledge of this protein, with particular attention to mutation-resistant residues or mutation "coldspots". In the present analysis, 54 mutation coldspots were identified and mostly were mapped to a dedicated function within the M-pro structure and function. Thus, the identification and short-listing of mutation coldspots of SARS-CoV-2 M-pro might single out mutation-resistant targets to be eventually inhibited, which indeed has two main advantages: (i) plausibly have a more potent effect on the protein inhibition and (ii) prevent mutation-based drug resistance.

Keywords: COVID-19, SARS-CoV-2, M-pro, 3CL-pro, genomic profiling, mutation coldspots, sequence analysis

1. INTRODUCTION

1.1. SARS-CoV-2 and the COVID-19 pandemic

SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) is the causative agent of the COVID-19 (COronaVIrus Disease 2019) pandemic [1]. After the COVID19 outbreak in December 2019, globally there have been 172,630,637 confirmed cases of this disease reported to the World Health Organization (see www.covid19.who.int), 3,718,683 of which have passed away (data updated on 6th June 2021). Therefore, COVID19 is a global threat to public health and human safety. COVID has a complicated pathogenesis ranging from mild fever, fatigue and dry cough to severe dyspnea and multiorgan failure in critical cases [1]. Upon binding of the spike (S) protein of SARS-CoV-2 to the human receptor, the angiotensin-converting enzyme 2 (hACE2), and entry to epithelial cells on the respiratory tract, viruses start replicating and migrating down to enter and infect alveolar epithelial cells in the lungs. This eventually triggers a solid immune response which may account for the pro-inflammatory phenotype and acute respiratory distress and failure [1].

1.1.1. SARS-CoV-2 phylogeny and genomic organization

Coronaviruses (CoVs) are a large group of enveloped positive-sense ssRNA viruses. Phylogenetically, they can be placed within the subfamily and family *Coronaviridae*, order *Nidovirales*. Four genera –namely, *Alphacoronavirus* (αCoV), *Betacoronavirus* (βCoV), *Gammacoronavirus* (γCoV) and *Deltacoronavirus* (δCoV) – make up this group [2]. Specifically, SARS-CoV-2 is clustered with βCoVs. It is known that while bats and rodents are the gene sources of most α- and β- CoVs, poultry and other birds have the same role in δ- and γ- CoVs [2]. CoVs have reiteratedly crossed species barriers. Prior to the outbreak of this novel CoV, only six CoVs –two αCoVs (*i.e.,* hCoV-229E and HKU-NL63) and four βCoVs (*i.e.,* hCoV-OC43, hCoV-HKU1, SARS-CoV and MERS-CoV)– had the ability to infect human. Among these, only SARS-CoV and MERS-CoV caused severe lower respiratory tract infections and extrapulmonary manifestations, which eventually result in death, similar to those reported in COVID19. On the other hand, the other human-infectious CoVs have self-limiting upper respiratory effects and occasional lower respiratory tract consequences in immunocompromised hosts and elderly [2]. Interestingly, both SARS-CoV and MERS-CoV originated from bats and both had an intermediate mammalian host –the Himalayan palm civet (*Paguma* larvata) and the dromedary camel (*Camelus dromedarius*) for SARS- and MERS- CoV, respectively– before finally jump to the human as a host. SARS-CoV-2 interspecies crossing ability suggests a similar transmission pattern to that found in SARS-CoV and MERS-CoV; that is, by means of an intermediate host. This host, however, is still to be elucidated. Nonetheless, several candidates, such as wild animals like the pangolin, the turtle, the snake or the ferret, as well as some domestic animals like cats, dogs, minks or even swine have been taken into consideration (reviewed elsewhere in detail in [3]).

The *Betacoronavirus* genus can be further subdivided in five subgenera: *Embecovirus*, *Merbecovirus*, *Nobecovirus*, *Hibecovirus* and *Sarbecovirus*, the latter containing SARS-CoV-2 (Figure 1). Phylogenetic analyses show that SARS-CoV-2 is found in a sister clade to the SARS-CoV and bat SARS-related-CoVs (SARSr-CoV) and the closest known neighbor is the bat coronavirus RaTG13 with an overall sequence identity of 96.2% [1,4]. Interestingly, although closely related, SARS-CoV-2 is clearly distinct to SARS-CoV and MERS-CoV having about 80% and 50% genomic similarity, respectively [5]. A more detailed analysis at an amino acid composition level show that generally SARS-CoV-2 is very similar to SARS-CoV, especially in some important proteins, such as the main protease (96% identity), the envelope protein (95% identity) or the nucleoprotein (94% identity). Nonetheless, other proteins such as the spike protein or the papain-like protein have less identity to their homologous in SARS-CoV with 76% identity in both cases [2].

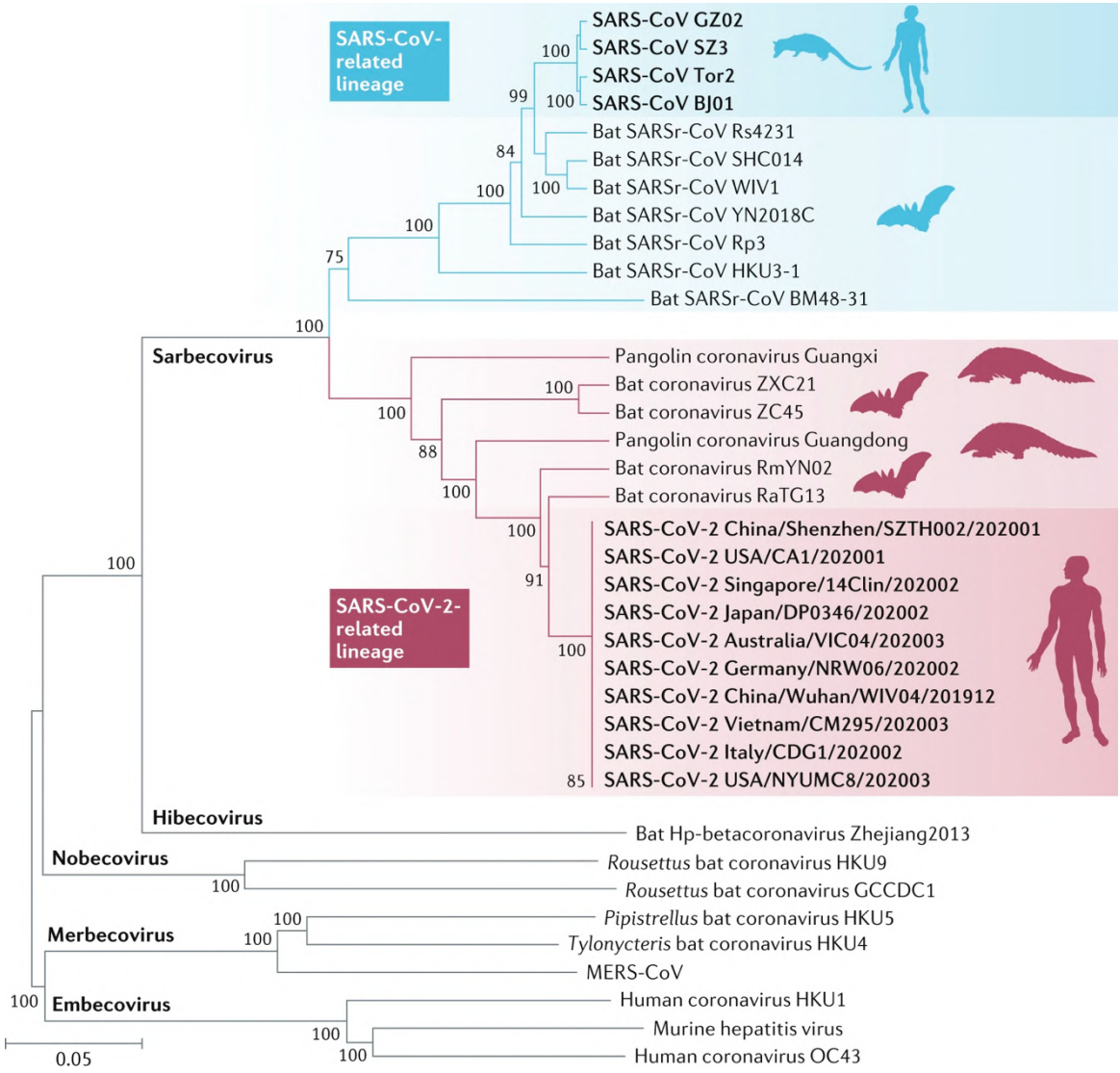

Figure 1 | Phylogenetic tree of full-length genome sequences of the β-CoV subgenus. Taken from Hu, B. *et al.* (2021).

SARS-CoV-2 has a very compact genome of about 29.9 kb in size encoding > 9,000 amino acids. The genetic makeup of SARS-CoV-2 contains two flanking untranslated regions (UTRs) composed by a 5' cap structure and a poly(A) 3' end, and 12 genes encoding 25 different proteins in between. The exact arrangement is as follows: 5' UTR–replicase ORF1ab–structural proteins[spike(S)–envelope(E)–membrane(M)–nucleocapsid (N)] – 3' poly(A). Other ORFs can be found within structural protein genes (Figure 2) [2,6]. Table 1 lists the exact function of all the encoded proteins as well as other interesting features. Interestingly, the SARS-CoV-2 RNA has an important secondary structure feature at the overlapping region between ORF1a and ORF1b: a -1 frameshift-stimulating pseudoknot. This structural characteristic enables a programmed -1 ribosomal frameshifting at genomic position 13,468 which is critical to produce essential proteins at a tightly regulated level. Indeed, it was found that complete inhibition of -1 programmed ribosomal frameshifting dramatically reduced SARS-CoV replication by some orders of magnitude. The full 3D structure of this pseudoknot has not been yet solved [7].

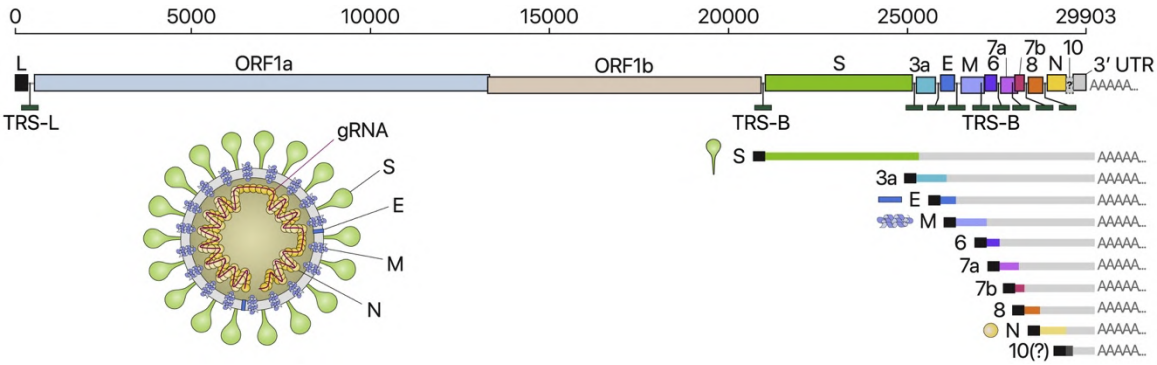

Figure 2 | Genomic organization of SARS-CoV-2. The exact position of each gene is depicted as well as a diagram to illustrate the function of structural proteins S, E, M and N. Taken from Kim *et al*., (2020), Cell 181, 914–921.

One interesting feature about SARS-CoV-2 –and other CoVs, in general– is the presence of two overlapping ORFs (*i.e.,* ORF1a and ORF1b) encoding for two polyproteins, pp1a and pp1ab, respectively (refer back to Figure 2). Polyproteins require an autoproteolytic cleavage to give rise to the 16 non-structural proteins that form the replicase/transcriptase complex (see Table 1). To this end, two proteases within this non-structural protein –namely, the main protease (M-pro) and the papain-like protease (PL-pro)– mediate this vital function. Table 1 shows the putative cleavage sites for each non-structural proteins within the polyprotein [1,8]. In fact, the M-pro is known to cleave no less than 11 peptidyl bonds on the large polyprotein 1ab including its own N-terminal and C-terminal processing sites [9,10]. Therefore, regions composing ORF1ab that encode different proteins are usually considered as distinct entities, resulting in 25 different "genes" in the genome.

It is worth mentioning that, despite slight overall dissimilarity in genomic identity with SARS-CoV, neither gross genomic organization nor ORFs and non-structural proteins (nsps) remarkable differences were found between this two clearly distinct species [2] (see Table 1). The major distinction between SARS-CoV and SARS-CoV-2 is found in two proteins: the spike protein and ORF8 [2].

Table 1 | Summary of the SARS-CoV-2 proteome, functions and proteolytic cleavage.

Gene/Protein name	Alternative name	Genomic start	Genomic end	Protein length	Amino acid identity to SARS-CoV (%)[1]	Putative cleave site[1]	Function[2]
nsp1	Leader	266	805	180	84	(LNGG'AYTR)	Inhibit host gene expression and interferon signaling.
nsp2		806	2719	638	68	(LKGG'APTK)	Accessory protein which may assist other proteins. Mostly unknown.
nsp3	Papain-like protease; PL-pro	2720	8554	1945	76	(LKGG'KIVN)	Papain-like protease with phosphatase activity. Involved in proteolytic cleavage of the polyprotein and inhibition of NFκß and p53 signaling.
nsp4		8555	10054	500	80	(AVLQ'SGFR)	Essential to membrane rearrangements during replication.
nsp5	3C-like proteinase; Main protease; Mpro	10055	10972	306	96	(VTFQ'SAVK)	Crucial in proteolytic cleavage of the polyprotein to generate the active form of nonstructural proteins.
nsp6		10973	11842	290	88	(ATVQ'SKMS)	Membrane rearrangements and autophagy.
nsp7		11843	12091	83	99	(ATLQ'AIAS)	Part of the replication complex formed by nsp7-nsp12-nsp8. Forms a multimeric complex with nsp8 that serves a processivity clamp for the RNA-dependent RNA polymerase.
nsp8		12092	12685	198	97	(VKLQ'NNEL)	Analogous to nsp7.
nsp9		12686	13024	113	97	(VRLQ'AGNA)	Protect viral RNA from degradation during replication.
nsp10		13025	13441	139	97	(PMLQ'SADA)	Forms complex with nsp14 and nsp16.
nsp12	RNA-dependent RNA polymerase; RdRp	13442	16236	932	96	(TVLQ'AVGA)	RNA-dependent RNA polymerase, essential for the replication.
nsp13	Helicase	16237	18039	601	100	(ATLQ'AENV)	RNA helicase with NTPase, dNTPase and RTPase activities
nsp14	3'-5' exonuclease	18040	19620	527	95	(TRLQ'SLEN)	3'-5' exonuclease with proofreading activity
nsp15	endoRNAse	19621	20658	346	89	(PKLQ'SSQA)	poly(U)-specific endoribonuclease
nsp16	2' O-Ribose methyltransferase	20659	21552	298	93	(end of orf1b)	Formation of cap in viral RNA in complex with nsp10.
S	Spike	21563	25384	1273	76	n. a.	Glycosylated protein known to mediate viral cell-host receptor interactions.
ORF3a		25393	26220	275	72	n. a.	Homotetramer formation, with ion channel properties. Linked to inflammatory, IFN and innate immunity responses and also to modulation of cell cycle.
E	Envelope	26245	26472	75	95	n. a.	Minor structural protein involved in the formation of channels in the ER of the host cell.
M	Membrane	26523	27191	222	91	n. a.	Membrane glycoprotein needed in membrane curvature, packing of RNA and budding of new particles.
ORF6		27202	27387	61	69	n. a.	Accessory protein involved in enhancement of viral replication.
ORF7a		27394	27759	121	85	n. a.	Binding to BTS-2 to prevent virus tethering on plasma.
ORF7b		27756	27887	43	81	n. a.	Integral transmembrane protein. Mostly unknown.
ORF8		27894	28259	121		n. a.	Accessory protein involved in enhancement of viral replication.
N	Nucleocapsid	28274	29533	419	94	n. a.	RNA packing to from a ribonucleocapsid. Critical role in viral assembly.
ORF10		29558	29674	38	-	n. a.	Accessory protein, probably linked to inhibiting the ubiquitin-proteasome system.

[1] Taken from: Chan *et al.* (2020). [2] Taken from: Prates *et al.* (2021). n.a.: not applicable

The spike protein (S) is a glycosylated protein that mediates the obligated interaction between the host and the virus to enable the entry to the host cell. Upon the entry, the viral RNA is ready to be translated. It is a 1273 aa protein that consists of an N- extracellular domain, a transmembrane (TM) domain and a small intracellular fraction [11]. The S1 subunit comprises the N-terminal domain and the receptor binding domain (RBD) while the S2 subunit is composed by the fusion peptide, two consecutive heptapeptide repeat sequences, TM domain and cytoplasm domain [11]. Interestingly, the S2 subunit is highly conserved, thereby being S1 the responsible for the most noticeable differences. Nonetheless, the core domain of RBD, the receptor-binding motif, is highly conserved and only external residues have been mutated. Strikingly, these residues have been reported as key substitutions in enhancing SARS-CoV-2 interactions with the ACE2 receptor [2,11].

On the other hand, ORF8, which has a role in enhancing viral replication, has significant changes between SARS-CoV and SARS-CoV-2. For instance, SARS-CoV-2 is depleted from an aggregation motif VLVVL found in all SARS-CoVs, which has been linked to intracellular stress pathways and the activation of inflammasomes [2].

1.2. SARS-CoV-2 main protease (M-pro)

As just stated, SARS-CoV-2 main protease (M-pro), also known as 3C-like proteinase, is crucial for the virus replication. Moreover, as Zhang *et al.* [10] stated, no homologous protein in humans are known, which makes the inhibition of this target even more interesting because inhibitors are unlikely to be toxic. Given this importance it appears to be reasonable that M-pro is being intensively studied as a pharmacological target against COVID-19, specially by computational methods [12]. For instance, Gimeno *et al.* [13], from the Cheminformatics and Nutrition research group, reported seven possible inhibitors of the SARS-CoV-2 M-pro which were predicted through consensus docking and are approved drugs for other uses. Among these inhibitors they reported that at 50 μM Carprofen and Celecoxib inhibit the main protease *in vitro* by 3.97% and 11.90%. More recently, another drug found within these inhibitors, Sarafloxacin, has been proved to show a 20.00% M-pro inhibition at 50 μM in the same assay (data not published).

SARS-CoV-2 M-pro is a 306-residue length chymotrypsin-like protease. Three domains can be identified within the protein: domain I (residues 8-101), domain II (residues 102-184) and domain III (residues 201-303) [9,14] (Figure 3A). The enzyme binding site is located at a cleft present between domains I and II. Domain III, in contrast, is a globular structure composed by five helices involved in regulating dimerization, as M-pro is known to be catalytically inactive as a monomer [10]. Thus, domain III is not directly involved in M-pro catalytic activity but rather appears to be indispensable for M-pro function *in vivo*. In fact, dimerization is mediated by interactions between domain II of one monomer and the N-terminal region ("N-finger", residues 1-7) of the other chain in a contact interface of ~1,394 $Å^2$. Molecular dynamics (MD) simulations showed that the domain I/II arrangements appear to be

pretty stable during all simulations, while domain III underwent ample reorientation in a monomeric form [15]. Interestingly, the N-finger must squeeze in between domains II and III of the monomer and domain III of the opposing chain to reach an important residue, Glu166, which is buried in the domain II/III interface. By doing this, the N-finger helps in binding site shaping [10].

Unlike other chymotrypsin-like proteases, M-pro has a non-canonical catalytic dyad composed by His41 and Cys145 instead of the usual Ser/Cys-His-Asp/Glu triad [14] (Figure 3B and C). Four different important subsites can be identified within the binding site: S1, S1', S2 and S3. Despite some differences in the exact composition in recent publications, the SARS-CoV-2 M-pro binindg site is formed by the residues depicted in Figure 3B [13]. S1 is composed by Phe140, Asn142, His163, Glu166 and His172 side chains with a contribution of the main chains Phe140 and Leu141. The S1' subsite, is formed by three different threonine residues (residues 24-26) which interact with the substrate mainly via H-bonds. S2 is a hydrophobic cleft laterally defined by Val168, Asp187 and Arg188 main chains, the side chain of the catalytic His41, Asp178 and Gln189 and Met165 at the floor. SARS-CoV and SARS-CoV-2 S2 (~252 Å^3) is indeed larger and presents a different shape than other CoV homologues

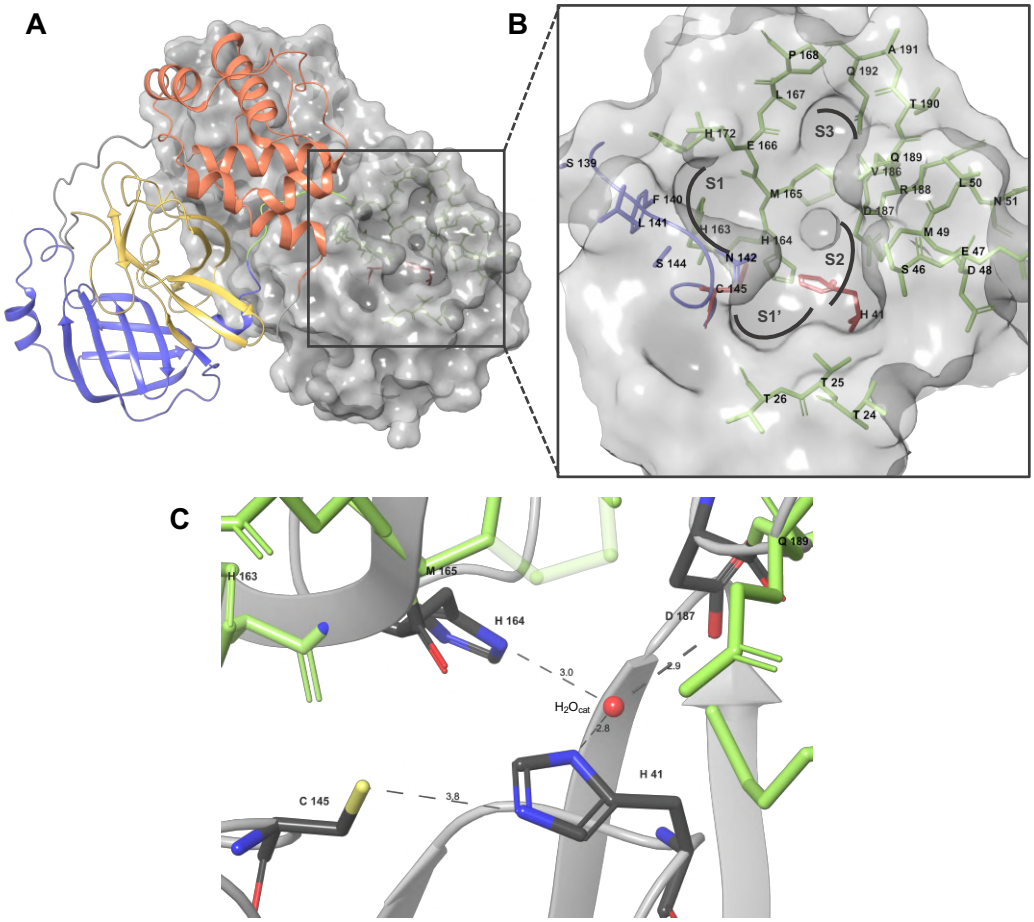

Figure 3 | SARS-CoV-2 main protease structure. Panel **A** shows the biological assembly of the M-pro in its dimeric form. One protomer (left) is shown in cartoon representation highlighting the different regions within the M-pro: N-finger (light green), domain I (blue), domain II (yellow), domain III (orange) and regions between domains (gray). The other protomer (right) is displayed as a surface with the residues highlighted. Panel **B** shows a more detailed view of the M-pro binding site. Residues of the binding site are displayed as backbone structures in different colors: residues in the binding site are displayed in green and the catalytic dyad residues and the residues composing the oxyanion loop are shown in red and blue, respectively. The exposed areas of different subsites are specified with a gray line. Panel **C** shows a detailed snapshot of the catalytic dyad and other residues interacting with H_2O_{cat}, which are colored in CPK. Distances between specific atoms are also shown, in angstroms (Å). This figure has been generated using Schrödinger Maestro and using the structure of SARS-CoV-2 free protein with PDBid code 6WQF.

of the α-CoV genus (*e.g.,* hCoV-NL63) which show a smaller S2 (~34 Å^3). S3 is defined by the flexible loop involving residues 165-168 and 189-192. This subsite is more superficial than S1 and S2 and it goes through extensive rearrangement upon ligand binding [12].

Structurally, Cys145 and His41 reactive atoms (Sγ and Nε2, respectively) are located pretty far within the binding site, at a distance of 3.8Å. Cys145 is part of the oxyanion loop, an S-shaped loop of domain II formed by residues Gly138-Gly146. Moreover, amide nitrogen of Gly143 and Cys145 define the "oxyanion hole", which will bind to the scissile peptide bond of the substrate. His41, which is supposed to act as a base during the nucleophilic attack or simply triggered by substrate binding, belongs to a small helix in domain I [14,15]. A water molecule, H_2O_{cat}, appears to play an important paper within the binding site, probably acting as the missing residue of a canonical binding site [14]. For instance, H_2O_{cat} enables the correct positioning of His41 through a complex H-bond network involving also Asp187 and His164 (Figure 3C) [14,15]. It is worth mentioning that the structure and composition of the binding sites greatly restricts the composition of the cleave sites that can be processed by M-pro, although the enzyme shows sequence promiscuity [14]. The S1 subsite has a doubtless requirement for a Gln, while small residues such as Ser and Ala are preferred in the S1' subsite. The deep hydrophobic S2 subsite has a preference for hydrophobic residues (*e.g.*, Leu, Phe, Val) (see Table 1) [9,12,16].

The reaction mechanism of SARS-CoV-2 M-pro is based on a common nucleophilic-type reaction (Figure 4). Mechanistically, His41 acts as a general base and there is a nucleophilic attack of the Cys145-Sγ to the peptide bond between P1 and P1' (*i.e.*, between Gln and Ser/Ala). This step is called the acylation step. In the second step, the deacylation step, the covalent bond to Cys154 is broken, so that the C-terminus of the excised (P fragment) is released. After forming a Michaelis complex with the residues of the catalytic center, during the first step of the reaction the peptide is excised at the preferred Gln (P1) residue and the thiol group of the catalytic Cys145 is acylated with the C-terminus of the excised peptide while the N-terminus (P' fragment) is released. During the first step, His41 Nε activates the thiol group of Cys145 by taking a proton from it and then transferring it to N (P1') followed by the nucleophilic attack of Cys145. In the deacylation step, the covalent bond to Cys145 is hydrolyzed by a water molecule activated by His41 and the enzyme is regenerated after transference of a proton from His41 to Cys145 [17,18].

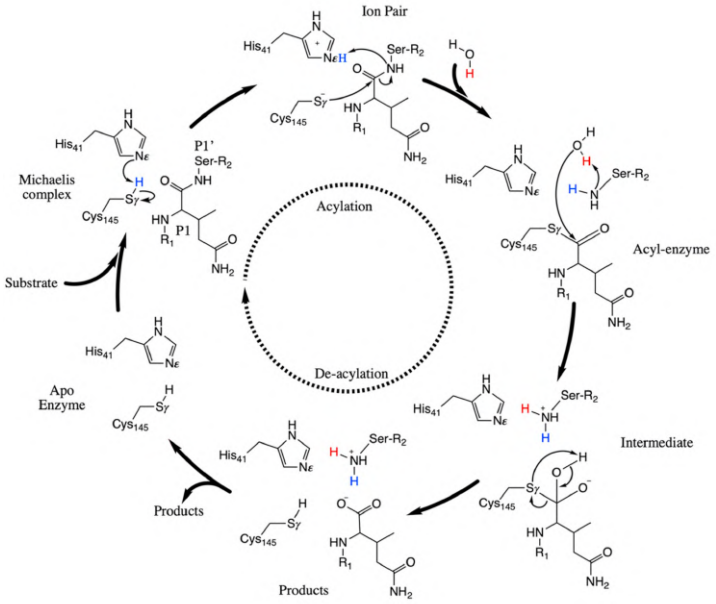

Figure 4 | Catalytic mechanism of SARS-CoV-2 M-pro. Taken from: Ramos-Guzmán *et al.* (2020).

SARS-CoV2 and SARS-CoV M-pro are very similar (see Table 1), with a 96% of identity. Only 12 amino acids differ from SARS-CoV and SARS-CoV-2 M-pros: Thr35Val, Ala46Ser, Ser65Asn, Leu86Val, Arg88Lys, Ser94Ala, His134Phe, Lys180Asn, Leu202Val, Ala267Ser, Thr285Ala, Ile286Leu, respectively. Moreover, this high identity is also represented in structural terms, with an RMSD value between the backbone structure of 0.884 Å [19]. Nonetheless, Gahlawat *et al.* [19] reported significant differences between the active sites of these two proteins. For instance, the binding site of SARS-CoV-2 M-pro was reported to be 319Å^3 whereas the homologous protein for SARS-CoV had a binding site volume of 224Å. Interestingly, this was linked to the presence of a mutation, Ala46 to Ser46 from SARS-CoV to SARS-CoV-2 M-pro. This Ser residue, located in the entrance of the binding site, enables H-bond interactions with two adjacent Thr residues (*i.e.,* Thr24 and Thr45), which are reported to be part of the binding site in SARS-CoV-2 M-pro but not in the SARS-CoV protease. Moreover, it makes the binding site entrance more hydrophilic [19]. Another interesting mutation is His134Phe. The phenylalanine residue present in the SARS-CoV-2 M-pro does not allow an H-bond interaction between the original His134 and Pro132, thus leaving freer the oxyanion loop. This lack of interaction makes the catalytic dyad to meet closer and could result in a faster transference of a proton between them [19]. Interestingly, the Thr285Ala, along with Ile286Leu mutation and the conserved Ser284, also have an important role in the catalytic efficiency. Indeed, mutating these three residues to Ala in SARS-CoV M-pro leads to a 3.6-fold enhancement of the catalytic activity of the protease concomitant with a tighter packing of the two monomers. Thus, the two naturally-occurring mutations in SARS-CoV-2 M-pro (*i.e.,* Thr285Ala and Ile286Leu) might account for the slightly higher catalytic efficiency of SARS-CoV-2 M-pro [10]. MD simulations also identified Ala285/Leu286 as a crucial interprotomer spot [15], which was clearly affected by ligand binding.

As previously reported, 11 sites are cleaved by M-pro, while the rest three are processed by the papain-like protease. Interestingly, the papain-like protease appears to cleave the polyprotein before M-pro is released [20]. M-pro is able to autoprocess itself and become mature in an infected cell. The exact mechanism by which this happens, however, is still to be clearly elucidated. Muramatsu *et al.* [20] reported a consistent model for the autoprocessing of SARS-CoV main protease. It is likely that the pro-dimer of the protease cleaves the monomeric form of the pro-form in a *trans* manner. Thus, pro-forms are plausible to dimerize. Interestingly, there is a first preferred cleavage of a monomer's N-terminal region to continue with the same region of the other monomer. Then, it is thought to be a similar rearrangement of the protomers to that one that occurs in the mature dimeric enzyme an both C-termini are excised. Nonetheless, it is worth reiterating that the exact autoprocessing of either SARS-CoV or SARS-CoV-2 M-pros is not known yet.

2. HYPOTHESIS AND OBJECTIVES

Since the outbreak of the COVID-19 pandemic, an unprecedented bunch of work has emerged in this field in part as a result of the scientific effort to obtain an effective cure and/or vaccines. It is of vital importance for these treatments to be useful in the mid-long term that their targets remain unaltered throughout the virus evolution. At the same time, the detection of SARS-CoV-2 infection, which usually relies on quantitative PCR analysis, requires binding of primers to specific regions within the genome. In this sense, it is crucial to keep track of its genomic variants.

Until 31st December 2020, more than two hundred thousand SARS-CoV-2 genomes were published in the Global Initiative on Sharing Avian Influenza Data (GISAID) (www.gisaid.org) and the number is still growing exponentially. By integrating this information, the mutational profile of SARS-CoV-2 could be obtained and used to shed light on the current knowledge of virus infectivity basis. This includes characterization of frequent mutations, more or less mutated genes and distinct regions in the genome with different mutational loads (*i.e.*, hypervariable and conserved regions).

On the other hand, mutational coldspots (*i.e.*, mutation-resistant residues) may correspond to important residues for a protein –at a structural and functional level– and, therefore, the identification could be useful to define putative binding sites for antivirals. To this end, given the key role of M-pro in SARS-CoV-2 replication, it might be interesting to define mutational coldspots in this target.

The main objectives of the present work can be summarized as follows:

1. Characterize the gross mutational profile of SARS-CoV-2 genome.

2. Report mutational hotspots and coldspots in the SARS-CoV-2 M-pro.

3. Understand the biological importance of the defined mutational coldspots of the SARS-CoV-2 M-pro.

3. MATERIALS AND METHODS

3.1. Sequence retrieval and pairwise alignment

274,504 complete full-length SARS-CoV-2 genomic sequences available on date 31st December 2020 were downloaded from the Global Initiative on Sharing Avian Influenza Data (GISAID) (www.gisaid.org) on 2nd February 2021. Before any step, some filters were applied in order to guarantee the quality of the sequences: (i) consider only sequences obtained from samples extracted from humans, (ii) avoid considering partial sequences by only keeping sequences with a minimum length of 29.000 bp and (iii) remove sequences not labeled as "high coverage" (*i.e.,* sequences containing: (i) less than 1 % unidentified bases (Ns), (ii) less than 0.05% of unique amino acid mutations, to withdraw possible sequencing artifacts, and (iii) no insertions and/or deletions, unless verified by the submitter).

The remaining 269,075 sequences were subjected to pairwise alignment to a reference genome. The complete genome NC_045512.2, isolated from Wuhan and submitted to the GenBank (https://www.ncbi.nlm.nih.gov/genbank/) database on 17th January 2020 was used as the reference genome. Thus, all coding regions of each sequence were aligned to the respective regions of the reference genome using the blastn algorithm. From this pairwise analysis two files containing information of all analyzed sequences were obtained. It is important to mention that only single nucleotide substitutions, but not insertions and deletions, are considered. Also, note that all available RNA sequences deposited in GISAID have thymidine bases instead of uracil bases, which are indeed the appropriate base for RNAs. This is a standard practice in sequence analysis, but it is worth to mention that all Ts in the present study will refer to Us in an eventual original RNA sequence.

File 1 is a 35609-rows-x-7-column data frame which contains information about each different mutation found. In this file, each observation (*i.e.,* each row) is a distinct mutation and each column adds some information about the mutation. For instance, for the first row, the mutation A23403G (*i.e.,* A in position 23,403 was mutated to G), that this mutation (see column "n_found") was found 253,060 times in the sequences considered and the first time it was found was 24th January 2020 (see columns "n_found" and "date_first_found", respectively). More information, such as the amino acid affected, whether the mutation is synonymous and the countries on which the mutation was found is also available in File 1. Figure 5 shows an extract from File 1.

mutation	gene	position	codon	aa	n_found	n_countries	date_first_found	id_first_found	countries	codon_position	is_synonymous	n_found_x100/n
A23403G	spike	23403	GAT614GGT	D614G	253060	140	1/24/20	EPI_ISL_422425	Germany; Brazil; Mexico; Italy; China; Switz	2	0	93.9399
C3037T	nsp3	3037	TTC106TTT	F106F	252778	140	1/24/20	EPI_ISL_422425	Germany; Brazil; Mexico; Italy; Switzerland;	3	1	93.8352
C14408T	RNA_pol	14408	CCT323CTT	P323L	252698	139	1/24/20	EPI_ISL_422425	Germany; Brazil; Mexico; Italy; Switzerland;	2	0	93.8055
G28881A	N	28881	AGG203AAA	R203K	81230	111	2/16/20	EPI_ISL_466615	Germany; Mexico; Switzerland; UK; Netherl	2	0	30.1539
G28882A	N	28882	AGG203AAA	R203K	80984	111	2/16/20	EPI_ISL_466615	Germany; Mexico; Switzerland; UK; Netherl	3	0	30.0625
G28883C	N	28883	GGA204CGA	G204R	80954	111	2/16/20	EPI_ISL_466615	Germany; Mexico; Switzerland; UK; Netherl	1	0	30.0514
C22227T	spike	22227	GCT222GTT	A222V	67911	49	3/13/20	EPI_ISL_467121	UK; Usa; Greece; Spain; Senegal; South_afr	2	0	25.2096
C6286T	nsp3	6286	ACC1189ACT	T1189T	67604	48	3/4/20	EPI_ISL_653254	Spain; Usa; UK; China; Switzerland; Romani	3	1	25.0957
G29645T	ORF10	29645	GTA30TTA	V30L	67492	47	3/11/20	EPI_ISL_444822	Drc; Denmark; Israel; India; UK; Switzerland	1	0	25.0541
C28932T	N	28932	GCT220GTT	A220V	67401	45	3/16/20	EPI_ISL_699657	UK; Netherlands; Senegal; Switzerland; Port	2	0	25.0203
G21255C	methyltransferase	21255	GCG199GCC	A199A	67355	45	3/13/20	EPI_ISL_416709	Usa; UK; Switzerland; Bangladesh; Australia	3	1	25.0032
T445C	leader	445	GTT60GTC	V60V	67051	41	3/16/20	EPI_ISL_699657	Switzerland; Ireland; UK; Hong_kong; Nethe	3	1	24.8904
C26801G	M	26801	CTC93CTG	L93L	66720	42	3/15/20	EPI_ISL_535802	UK; Iceland; Usa; Spain; Switzerland; Irelan	3	1	24.7675
G25563T	ORF3a	25563	CAG57CAT	Q57H	62166	115	2/3/20	EPI_ISL_489996	Netherlands; Taiwan; Usa; France; Finland;	3	0	23.077

Figure 5 | Extract from File 1

File 2 is a 269,075 x 161 matrix that contains all the genomes being analyzed. The exact mutations found in each gene in each genome are listed here, along with other information about the genome (*e.g.,* country/city where the genome was sequenced, the length of the genome, information about the submitter, among others).

3.2. SARS-CoV-2 M-pro sequence alignment with other coronavirus and mutations analysis

All the herein performed representations and statistical analyses are performed using the programming language R [21]. Generally, representations are made using R package `ggplot2` [22].

For searching main proteases of different coronavirus used in the multiple sequence alignment (Figure 16), structures with at least 90% sequence similarity with SARS-CoV-2 M-pro were obtained from the Protein Data Bank (PDB, www.rcsb.org) and those corresponding to SARS-CoV-2 were not considered. The SARS-CoV-2 M-pro sequence and structure is that with PDBid 6WQF. Then, sequences were aligned with R package `msa` [23] using ClustalW with the default settings as the alignment algorithm. Also, the function `msaPrettyPrint()` was used to customize the plot of multiple sequences alignment using tailored LaTeX code. The secondary structure assigned by STRIDE [24] to the PDBid 6WQF was also added to Figure 16 using `msaPrettyPrint().`

Mutational coldspots have been computationally annotated considering those codons that: (i) do not have reported mutations in any of the three genomic positions composing the codon (*i.e.,* conserved codon) or (ii) despite having mutations, all mutations within the codon are synonymous (*i.e.,* synonymous or missense mutation-resistant codon).

4. RESULTS AND DISCUSSION

4.1. Mutational profile of the SARS-CoV-2 genome

The results from the present work are based on the analysis of 269,075 full-length genomic sequences available in GISAID (www.gisaid.org) from December 2019 to 31st of December 2020 (see Materials and Methods section for further details). It is worth bearing in mind that there is a bias in the genomes being analyzed, as the sequencing rates in different countries greatly vary. This is indeed important to manage conclusions drawn from the present work, which will also reflect the bias present. For instance, the majority of the genomes (68%) were deposited only from 3 countries: UK, USA and Denmark. Table 2 shows the top 10 countries contributing to the analyzed genomes dataset. Nonetheless, under no circumstances this bias invalidates the results herein reported.

Table 2 | Top countries with higher contribution to the genomes present in the dataset

Country	Number of genomes
UK	107,285
Usa	56,728
Denmark	20,843
Australia	12,909
Japan	6,809
Canada	4,956
Netherlands	4,763
Switzerland	4,325
Spain	4,114
India	3,495

This mutational profiling of SARS-CoV-2 full-length genomes reported 3,520,709 total single nucleotide substitutions (SNSs), which brings about a mean number of 13.08 mutations per genome. Considering the length of the reference sequence NC_045512.2 (*i.e.,* 29,903 nucleotides), the mutation rate for SARS-CoV-2 equals to an estimated mutation rate mean of $4.4 \cdot 10^{-4}$ mutations per site per year. This mutation rate is slightly higher than that reported for SARS-CoV and MERS-CoV [25]. Viral mutation rates vary widely, specially due to the differences in the fidelity of the polymerases used in replication [26]. Among them, RNA viruses which use RNA-dependent RNA-polymerases (RdRp) usually have higher mutation rates, because RdRps are more prone to errors than RNA-rependent DNA-polymerases or Reverse transcriptase [26]. Nonetheless, it is interesting how coronavirus possess a unique feature: mutation rates are significantly lower than those reported in other RNA viruses, presumably due to nsp14's 3'-to-5' proofreading activity [25]. Also, this could explain the abnormally large genome observed in coronaviruses (from 27 to 32 kb), compared to usual ssRNA(+) viral genomes. From a simplistic point of view, it is generally assumed that larger genomes might have higher mutation rates. Thus, proofreading activity provided by nsp14 might make a balance between genome mutation –which indeed favors viral fitness– and correction of some mutations, to avoid the outnumbering of viable virions by unviable virions which had acquired too many mutations [27].

It is worth mentioning that of the total number of mutations reported in all genomes, some of them were repeated. Thus, only 35,609 unique mutations were annotated. Among them, transitions (*i.e.,* purine to purine or pyrimidine to pyrimidine) were more frequent (57.6 %) than transversions (*i.e.,* purine to pyrimidine, or vice versa) (42.4 %). A more detailed analysis of SNSs (Figure 6) shows a preference of T>C, A>G and C>T, that represent 17.4 %, 16.8 % and 12.5 % of unique mutations, respectively. Although these three mutations were also the most prominent ones in early stages of the pandemic, a rearrangement of their prevalence has been observed so that C>T is no longer the most prominent mutation [28]. As reported by Wang *et al.* [28], C>T were the most important mutations (in number) during the firsts months of the pandemic. In fact, the frequency of C>T mutation has decreased over the weeks (Figure 7), because as the pandemic goes on it is more difficult to find C that have not been mutated. In contrast, this tendency is not observed in any of the other mutations (data not shown). This was hypothesized to be the result of host-specific RNA gene editing, through APOBEC conserved cytidine deaminases, which may account for this abnormally large number of C>T mutations. Also, it was proposed that the higher number of T>C could be the result of viral protective mechanisms against defective mutations [28]. Nonetheless, this hypothesis still needs to be proved.

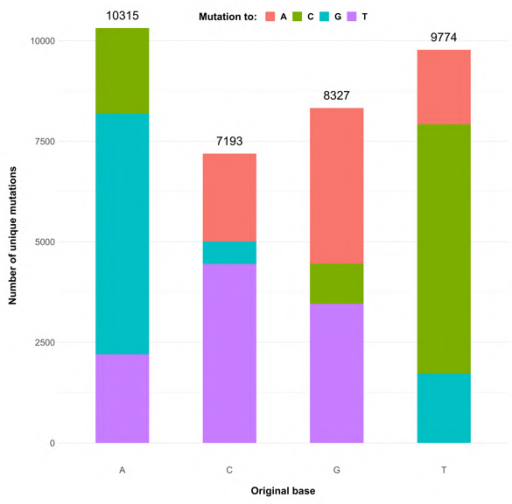

Figure 6 | Occurrences of SNS in the SARS-CoV-2 genome. The total length of each stacked bar represents the number of SNS from an original base in 35,609 unique mutations. Each portion represents the number of mutations to a specific base. The number above each bar equals to the number of unique mutations from the original base.

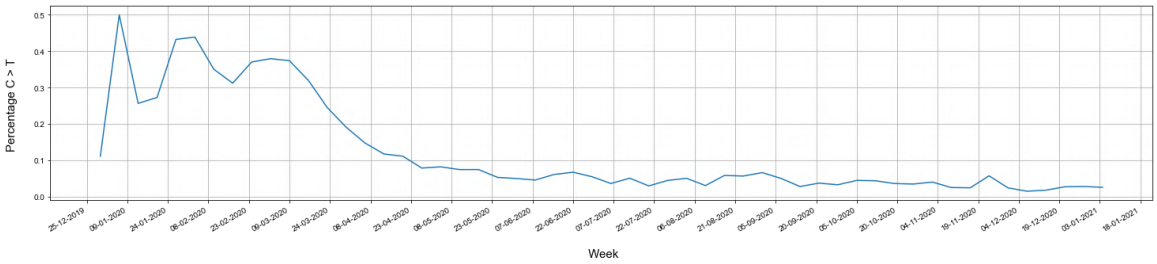

Figure 7 | Evolution of C>T mutation percentage of reported unique mutations over the weeks.

Figure 8 summarizes the results of gross mutational profiling of SARS-CoV-2 herein reported. As stated previously, there are mutations that have been found more than once and are more frequent than others (Figure 8A). For instance, the C3037T, C14408T and A23403G mutations from the nsp3, RdRp and spike genes, respectively, are in more than 93% of the genomes analyzed. These mutations are in more than 139 countries and they first appeared at the end of January. Among these three mutations, two of them, C14408T and A23403G represent an amino acid substitution in the RdRp gene and the spike gene and special attention has been given to the resulting variants. C14408T

mutation results in a substitution from a proline to a leucine in residue 282 of RdRp (P282L). This mutation, which falls out of the binding site of RdRp is located in an exposed surface of the enzyme [29]. Interestingly, Chand *et al.* [30] reported that such mutation might influence both the secondary and tertiary structure of RdRp, providing extra rigidity to the enzyme. However, since it is located in a poorly characterized region, the exact function of this favored substitution is not completely understood. On the other hand, the well-characterized A23403G mutation, which leads to D614G substitution in the spike protein is known to be linked to an increase in infectivity [31–33]. In fact, this mutation has superseded the original Asp614 and it is well extended worldwide. This is due to the fact that this mutation shifts the equilibrium of the spike protein towards a fusion-competent state by changing the conformation of the S1 domain [33]. Also, it was proven that this mutation improves spike incorporation in the virion [31], although other authors have not observed significant differences [33].

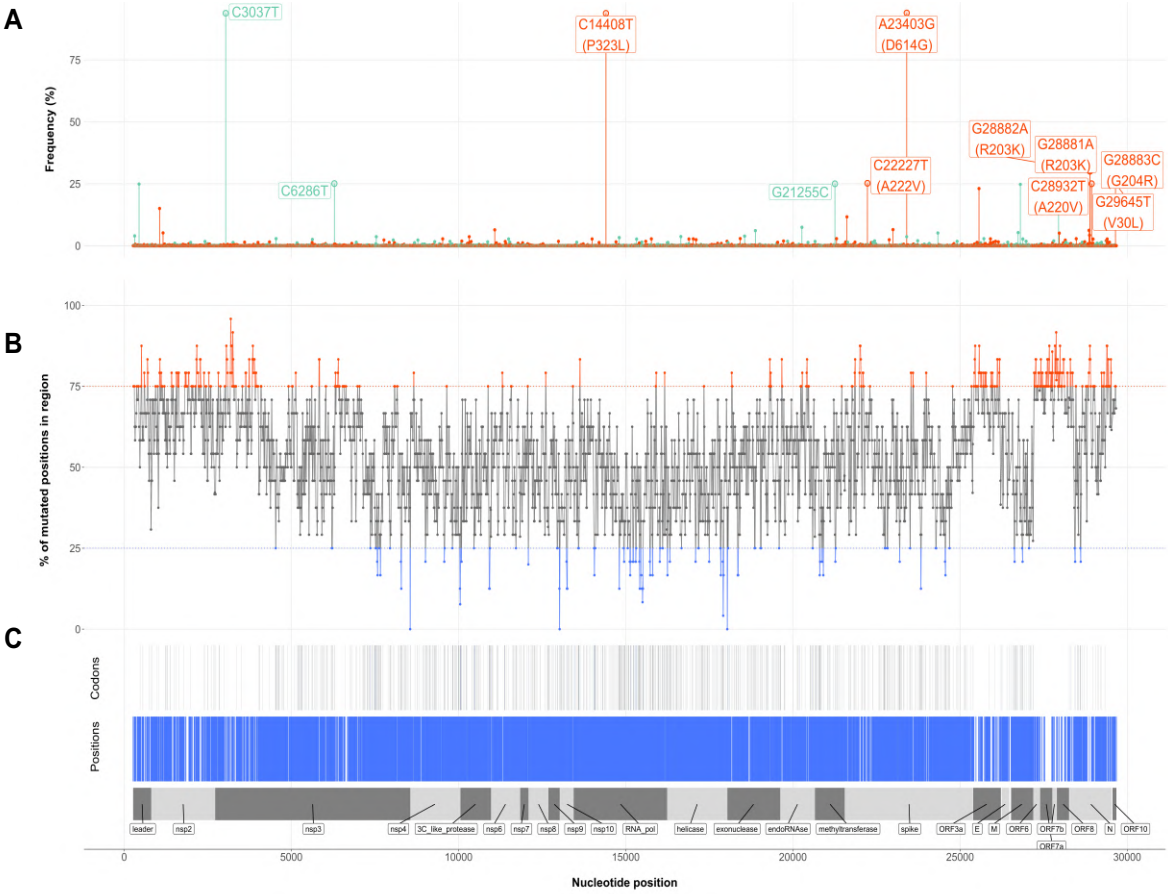

Figure 8 | Mutational profiling of SARS-CoV-2 M-pro genome. Panel **A** shows the reported ~35,000 unique mutations in the SARS-CoV-2 genome in the specific genomic position (x axis) and the frequency in which they are found (y axis). Note that all mutations found with a frequency (calculated as number of this specific mutation between total number of genomes) above 25.0% have a label with information about the mutation and the amino acid change between brackets (if applicable). Panel **B** shows an analysis of mutations in bins of the genome length (bin width, 24 nt; step width, 12 nt). Regions with more than 75 % sites mutated (at least 1 non-synonymous annotated) are highlighted in orange and are considered to be hypermutated regions. Analogously, regions with less than 25% of mutated nucleotides are highlighted in blue and are referred to as conserved regions. Panel **C** shows conserved codons and positions within the SARS-CoV-2 genome. A distinction is made between codons that have mutated but always in a silent way (gray) and truly conserved codons (blue). Conserved positions are displayed in blue.

The analysis of genomic region showed a clustering of hypermutated regions in both ends of the genome, while central regions were mostly conserved (Figure 8B). This brings about an obvious result, central regions of the genome, which encode proteins with a key role in viral replication, are indeed conserved. Apart from the exact conservation of each gene (see Figure 9) it is also interesting to consider the conservation of the sequence of an important structural feature of SARS-CoV-2 genome: the -1 frameshifting-stimulating pseudoknot [34]. In fact, the ribosomal frameshifting signal can be divided in three different structures: the "slippery site", a linker region and the frameshifting-stimulating pseudoknot. The "slippery site" spans genomic positions 13,462 to 13,467 having the sequence TTTAAA and it is situated just before the base that is repeated (C in position 13,468). The conservation analysis of the sequence shows an 83.3% of conservation with only one mutation in position 13,465. Nonetheless, it is worth mentioning that such mutation has been found only once in the 269,075 analyzed genomes. This indeed suggests an important degree of conservation. Also, this slippery site has been found to be conserved in SARS-CoV [34]. On the other hand, for the frameshifting-stimulating pseudoknot (positions 13,475 to 13,541) the degree of conservation decreases to 53.7%. This might support the fact that despite being required and crucial for an effective viral replication –so it is not a very variable region–, several frameshifting-stimulating structures could exist and be functional –so a large degree of conservation is not mandatory–. These results still support the fact that the ribosomal frameshifting signal could be an interesting target to impede SARS-CoV-2 replication.

As Figure 8B and Figure 8C show, there are some important differences in the genetic conservation degree of SARS-CoV-2. Thus, focusing on the number of unique mutations we wanted to address the mutation rates of each gene individually, normalizing by the length of each gene. As mutation rates per nucleotide appeared to be small and could be difficult to understand, the mutation rates per a thousand nucleotides are presented (calculated as mutation rate per nucleotide in each gene · 1,000) (Figure 9).

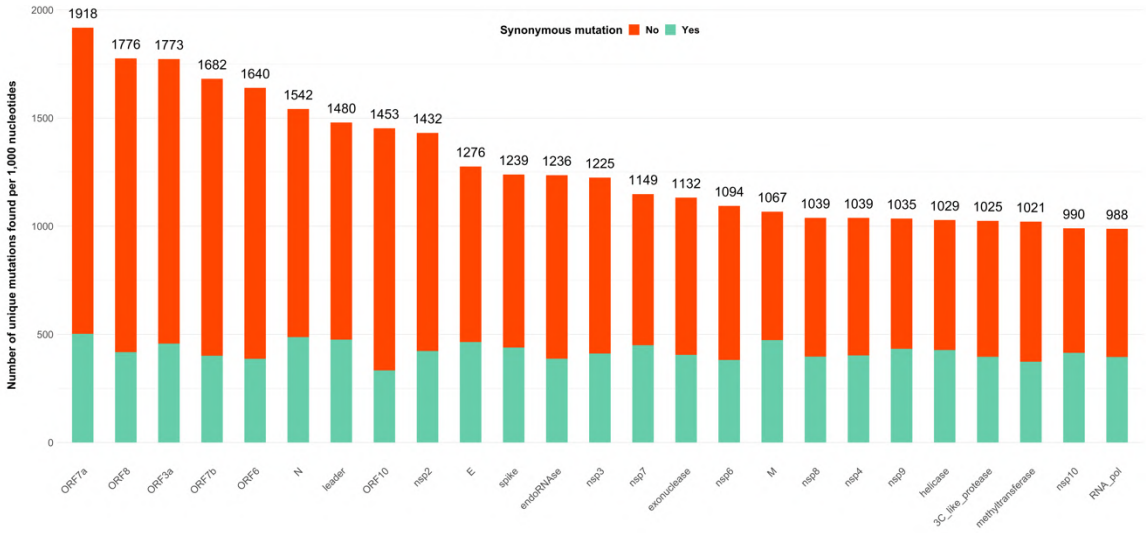

Figure 9 | Number of unique mutations per 1,000 nucleotides found in each of 25 genes of the SARS-CoV-2 genome. Mutation rates are calculated dividing the number of unique mutations of each gene by its length and multiplying by 1,000. Synonymous and non-synonymous mutations are colored in aquamarine blue and orange, respectively. The rounded numbers, without decimal positions, of mutations per 1,000 nucleotides are displayed on top of each bar.

Less mutations have been annotated for the genes that encode proteins that have a pivotal role in the replication of the virus (*e.g.,* helicase, M-pro, methyltransferase and RdRp) than for other genes with accessory functions (*e.g.,* OFR7a, ORF8, ORF3a, ORF7b and ORF6). Indeed, the number synonymous mutations per 1,000 nucleotides is similar in all genes, suggesting that non-deleterious mutations have a similar frequency in all genes and that there is not a clear preference of mutation in any gene. On the other hand, mutations that could affect to the protein function and/or structure (*i.e.,* non-synonymous) are selectively less found on genes encoding proteins that belong to the replicative machinery than on genes coding accessory proteins. This is also in line with previous reports from mid 2020, indicating that the tendency to conserve structural and functional important features is still maintained [35]. As reported by Jaroszewski *et al.* [35], it is important to consider differentially distinct parts of each gene. A good example of this is the N gene. In this case, regions of the gene coding for amino acids present in highly disordered structural domains are significantly more mutated than less flexible regions. In other words, well defined domains in the N protein carry less mutations than expected considering the total number of mutations in the gene. These results, albeit being truly interesting, require a meticulous analysis of the mutational profile in each gene, which completely falls out of the gross genomic characterization presented as the objective of this work.

4.2. Mutation hotspots in the SARS-CoV-2 M-pro

SARS-CoV-2 main protease is encoded by a highly conserved gene (refer back to Figure 9). Specifically, genomic analysis showed a total number of 63,633 genomes containing mutations, 38% of which were synonymous. Among the 63,633 genomes, only 941 unique mutations were identified. Considering the length of the M-pro gene (*i.e.,* 918 nucleotides), this gives a mutation ratio of 1.025 mutation per nucleotide.

Nonetheless, as previously discussed, it is worth mentioning that some genomic positions have reported more mutations than others (see Figure 8). For instance, mutation C10319T, which results in a Leu to Phe change in residue 89, is found in 3.7% of the analyzed genomes and represents roughly 15% of total 63,633 genomes carrying mutations found in M-pro gene (Figure 10B). In fact, altogether, top 5 mutations with higher occurrence frequencies shown in Figure 10B are found in up to 42% of genomes with mutations in the M-pro gene.

Interestingly, G10097A mutation defines a major clade in SARS-CoV-2 phylogeny, according to analyses carried out by the open-source project NextStrain [36] (available at: nextstrain.org/ncov/global). Thus, all sequences belonging to clade 20D, which presumably originated from clade 20B during March 2020, carry this non-synonymous mutation that results in a Gly15Ser mutation. Nonetheless, although this mutation represents a change in the polarity of the residue (from a non-polar to a polar residue) and there might be a predicted clash to Lys97, since it is located in the outer surface of the protein, no important changes might arise from this mutation [37].

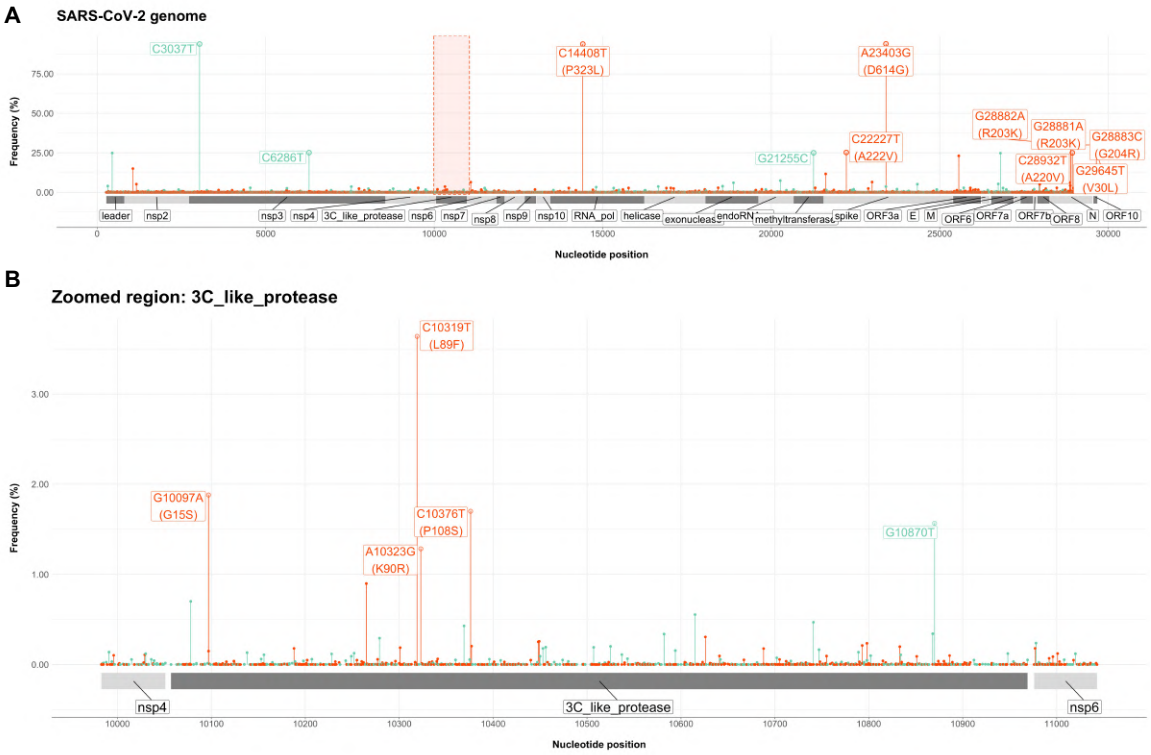

Figure 10 | Mutational profile of the SARS-CoV-2 genome with a zoom to the M-pro gene within the pp1ab polyprotein gene. Orange rectangle delineates the zoomed regions in Panel **B**. Panel **B** shows genomic mutations in the M-pro gene, and top five mutations are labelled just as in panel **A**. Synonymous mutation and non-synonymous mutations are depicted in aquamarine blue and orange, respectively.

Different is missense mutation A10323G, which results in K90R mutation on a protein level. Apart from being present in the lineage of the well-known "South-African variant", designated 20H/501Y.V2 or B.1.351, this mutation has been reported to increase the structural stability of the polyprotein pp1ab [38].

Changing the scope to a protein level, a mutational analysis of M-pro residues has been addressed (Figure 11). Beyond the aforementioned residues, the third more frequently mutated residue is Leu272 (Figure 11A), which is located in the outer surface of domain III. As the codons encoding Leu cannot mutate by changes in the 3rd position of the codon, although more than 5,000 genomes have been reported a mutation in this codon (being an important part the synonymous transversion G10870T (see Figure 10B), which indeed affects the third base of this codon), only 0.1% of these mutations are actually affecting the amino acid being encoded. Among the top 15 most frequently mutated residues, Asp176 has a unique feature: no missense mutations have been reported, thereby making Asp176 to be considered a mutation-resistant residue. The herein performed analysis reported 912 genomes on which Asp 176 appears to be mutated and they correspond to a unique mutation, C10582T, which appeared early in the COVID-19 pandemic, during February 2020. It is indeed curious that the only mutation reported in this codon, which is well extended worldwide (found in 38 different countries) is the only one that could not change the amino acid encoded. For instance, we could compare that to Ala191, which has a similar mutation frequency, with 833 individuals carrying mutations on it, but none of them are synonymous. Statistically, it is three times more probable that a single nucleotide

substitution leads to a synonymous mutation in an Ala residue than that would happen in an Asp residue (*i.e.,* 0.333 and 0.111 for Ala-to-Ala and Asp-to-Asp, respectively [39]), reinforcing the idea that Asp176 could be an important residue for M-pro structure and/or function.

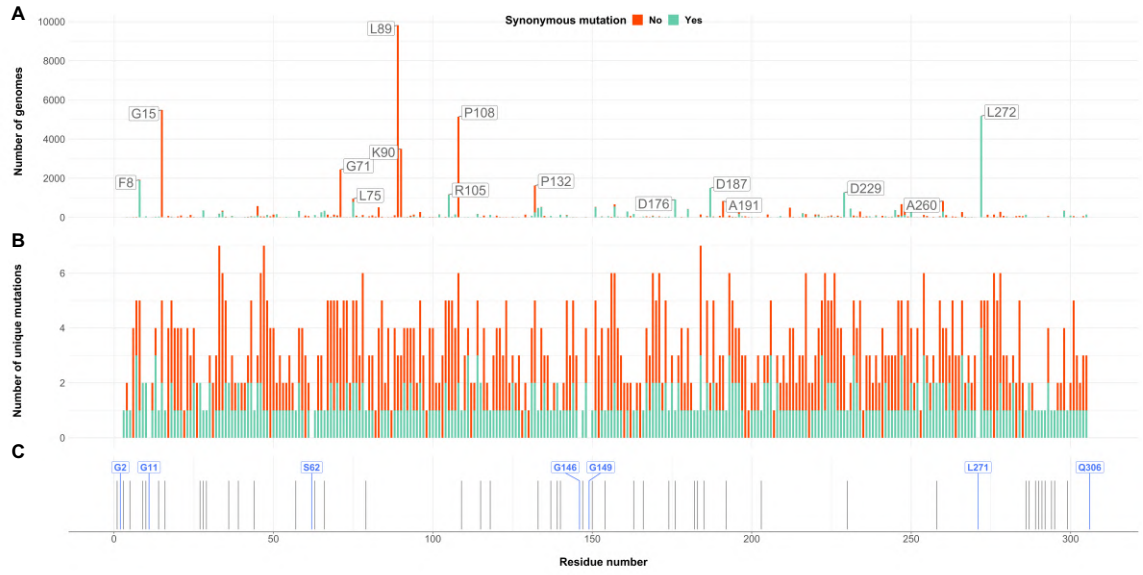

Figure 11 | Mutational profiling of the SARS-CoV-2 M-pro at a protein level. Mutations from each of the three codon positions are integrated and reported as mutations of the respective residues. Panel **A** shows the number of genomes (y axis) on which a mutation in a specific residue (x axis) was found. Top 15 residues found more frequently mutated are labeled. Panel **B** shows the number of distinct unique mutations (y axis) found in each residue (x axis). For both panels (**A** and **B**) a distinction is made between synonymous mutations (aquamarine blue) and non-synonymous mutations (orange). Panel **C** shows M-pro mutation coldspots and missense mutation-resistant residues and conserved codons are depicted in gray and blue, respectively. Conserved codons are also tagged with the residue and the position within the protein.

It is also interesting to consider the number of different mutations found in each residue of M-pro (Figure 11B). Although the number of achievable missense mutations depends on the encoded amino acid, evaluating different mutations in each residue could be useful in aiding to discern between fortuitous mutations with little or no effect to the protein and evolutionally selected mutations. For instance, mutations in Asp33 have been found in 186 genomes, among which 7 different mutations can be distinguished. This makes Asp33 the residue with more distinct mutations annotated hitherto. In contrast, Lys90, mutations of which have been found in 4,041 genomes, has only 3 different mutations –namely, K90R, K90N and K90K– the former representing roughly 85% of all findings. Considering that both Asp and Lys can be encoded by two different codons, the remarkable differences in the unique mutations between them, supports the idea that the effect of K90R mutation might have a positive effect and it is indeed evolutionally selected, although mutations on Lys90 have been found in up to 3,800 genomes more than mutations on Asp. Also, one could argue that this might happen because of differences in the first-found date, but in this specific example both mutations are dated during the first months of the COVID-19 pandemic. Nonetheless, it is important to note that such affirmations are only hypotheses which could be helpful in interpreting some of the mutational hotspots found here but, in any case, are unequivocable reported findings. Only one of the 15 residues which are mutated more frequently, Pro108, is found among the residues that have been reported to have more distinct mutations. Interestingly, although the probability of a Pro-to-Ser mutation is 3-fold

less than the Pro-to-Pro mutation by a single nucleotide substitution [39], P108S mutation is present in 4,579 genomes compared to 8 genomes of P108P mutation. That striking difference suggests a preferred Ser in the 108 position of the SARS-CoV-2 M-pro. Given that the residue in position 108 is highly exposed to solvent, mutation to a polar residue might be beneficial.

Another important result drawn from this mutational profile analysis concerns the M-pro binding site. The two residues composing the catalytic dyad (*i.e.,* His41 and Cys145) have reported mutations, although at very low frequencies. Two distinct mutations were found regarding His41, one being synonymous and the other being a missense mutation. Since the missense mutation, C10175T, was only detected once and that His41 mutation is very unlikely to be present in clinical isolates [40], because it will totally compromise viral replication, it could be assumed that His41 missense mutation is actually a sequencing artifact. On the other hand, three unique mutations were found on the catalytic Cys145. One of these mutations was synonymous whereas the latter two were missense mutations. Nonetheless, although they were found more than once, it is relevant that these mutations were found together in the same two genomes. Thus, G10488T and T10487A completely changed TGT codon encoding Cys145 for ATT codon, encoding an Ile residue. This mutation would affect M-pro function, as C145A mutants are unable to cleave peptide bonds [20], and it is indeed surprising to find such mutation two different clinical isolates. From Figure 11 it is clear that mutations in M-pro occur throughout the protein and that the binding site has a high tolerance to mutations, as previously reported [40]. In fact, from the present analysis, only 6 out of 32 residues in the binding site had either synonymous mutations or no reported mutations. Among these 6 mutation-resistant residues, only 1, Gly146, had not been mutated (Figure 11C). In fact, the binding site has a similar number of total mutations per residue than the complete protein, indicating again that the binding site of M-pro has certain tolerance to mutations (Table 3).

Table 3 | Number of unique mutations per residue in different regions of the M-pro protein

	Total unique mutations	Missense unique mutations	Synonymous unique mutations
N finger	2.000	1.000	1.000
Domain I	3.191	2.043	1.149
Domain II	2.976	1.807	1.169
Domain III	3.146	1.893	1.252
No domain	2.947	1.737	1.211
Binding site	3.000	1.906	1.094
Full protein	3.075	1.886	1.190

Interestingly, the most conserved region within M-pro is the N-finger, with an average of 2 mutations per residue. In fact, 4 out of 7 residues composing the N-finger (Ser1, Gly2, Phe3 and Lys5) are mutation-resistant residues, and specifically Gly2 is a mutational coldspot with no reported mutations. Residues Met6 and Ala7 accumulate 9 out of 14 mutations of the N-finger. The most frequently found mutations in these residues are M6L and A7V. In both cases, there is a little increase in the residue size

while preserving the hydrophobicity of the residue. Concerning A7V, Amamuddy *et al.* [41], have shown that such mutation increased the number of proximal contacts with neighboring residues and so did the hydrophobic interactions. Nonetheless, a predicted Van der Waals radius clash was also present when analyzing this mutation [41]. It is worth mentioning that, by means of MD simulations, Ala7 has shown a high rigidity [41]. Perhaps, given that A7V enables more hydrophobic interactions and proximal contacts, that mutation could be evolutionally selected since it might add rigidity to residue 7.

Other mutations, found earlier in the COVID-19 pandemic, are also reported here. For instance, the N274D mutation, which was found in 101 genomes, has been reported to induce significant changes within the binding site. Specifically, this mutation makes a non-canonical positioning of Phe140 possible, which flips to interact with residues from the C-terminal region of M-pro [41]. Such structural changes pull the oxyanion loop away from the catalytic dyad and make the binding site larger. Bzówka *et al.* [42] found a highly flexible loop in the vicinity of the binding site composed by residues 44-52, which has a potential role in regulating the access to the binding site. They also predicted the energetic favorableness of potential mutations of the whole M-pro protein. Evolution has indeed agreed with their predictions in some cases. For instance, mutations in Thr45 and Glu47 were predicted to be energetically favorable and they were reported whereas Ser39, on which mutations were not favorable, is a mutation-resistant residue. Also, they found that residues from the catalytic dyad were prone to mutate, and both reported mutations in the present analysis, as previously discussed.

4.3. Biological importance of SARS-CoV-2 M-pro coldspots and comparison to other CoVs

4.3.1. Structural implications of SARS-CoV-2 M-pro coldspots

About one year of mutations might be enough for the virus to accumulate some key mutations [43], but also to maintain residues which are important for its replication machinery. Recently, Krishnamoorty and Fakhro [44] reported a study of mutation coldspots of the M-pro. However, their work had two main limitations: data was downloaded in early November 2020, containing only 19,154 genomes with mutations in M-pro, and only missense mutations were considered. Here, given that an increase of the sequencing rate was found in mid-November (data not shown) and the importance of mutation-resistant residues (*i.e.,* residues that, if mutated, all mutations found are synonymous) apart from conserved residues, we wanted to further investigate the role of these presumably crucial residues for M-pro. For the sake of simplicity, from now on missense mutation-resistant residues and conserved residues will be referred as mutation coldspots, although the distinction between them will be made if necessary.

Table 4 lists all coldspots found in the present work (see also Figure 11C). Also, for missense mutation-resistant residues SNS mutations and their frequencies of mutation are displayed. As shown in Table 4, 54 mutation coldspots were identified in the present study.

Table 4 | Mutation coldspots of the SARS-CoV-2 M-pro

Residue	Codon sequence	SNS mutation	Frequency (%)
Ser1	AGT	T10057C	0.0011
Gly2	GGT	-	-
Phe3	TTT	T10063C	4e-04
Lys5	AAA	A10069G	0.0022
Pro9	CCA	A10081C; A10081G	0.0015; 0.0011
Ser10	TCT	T10084C; T10084G	0.0137; 0.0059
Gly11	GGT	-	-
Glu14	GAG	G10096A	0.0148
Cys16	TGT	T10102C	7e-04
Leu27	CTT	T10135A; T10135C	0.0033; 0.0015
Asn28	AAC	C10138T	0.1325
Gly29	GGT	T10141C	0.0011
Val36	GTT	T10162C; T10162A	7e-04; 4e-04
Pro39	CCA	A10171G; A10171C	0.0022; 7e-04
Cys44	TGC	C10186T	0.016
Leu57	TTA	A10225G	0.0041
Ser62	TCT	-	-
Asn63	AAT	T10243C	0.0434
Phe66	TTC	C10252T	0.1255
Gly79	GGA	A10291G	0.0045
Gly109	GGA	A10381G	0.0022
Leu115	TTA	A10399G; T10397C	0.0071; 7e-04
Tyr118	TAC	C10408T	0.0468
Asn133	AAT	T10453C	0.1782
Lys137	AAG	G10465A	0.0275
Ser139	TCA	A10471G; A10471C	0.0037; 4e-04
Phe140	TTC	C10474T	0.0509
Gly146	GGT	-	-
Ser147	AGT	T10495C	0.0022
Gly149	GGT	-	-
Phe150	TTT	T10504C	4e-04
Tyr154	TAT	T10516C	0.0011
His163	CAC	C10543T	0.0653
Glu166	GAA	A10552G	0.0145
Gly174	GGC	C10576T	0.0126
Asp176	GAC	C10582T	0.3385
Tyr182	TAT	T10600C	0.0015
Gly183	GGA	A10603G	0.0019
Phe185	TTT	T10609C	4e-04

Table 4 | *(cont.)*

Residue	Codon sequence	SNS mutation	Frequency (%)
Gln192	CAA	A10630G	7e-04
Asn203	AAT	T10663C	4e-04
Phe230	TTT	T10744C	4e-04
Gly258	GGA	A10828C; A10828G	0.0011; 4e-04
Leu271	TTA	-	-
Leu286	TTA	A10912G	0.0501
Leu287	TTA	T10913C; A10915G	0.0037; 4e-04
Asp289	GAT	T10921C	0.0052
Glu290	GAA	A10924G	0.0037
Phe291	TTT	T10927C	0.0022
Thr292	ACA	A10930G	0.0011
Phe294	TTT	T10936C	0.0015
Asp295	GAT	T10939C	0.0026
Gln299	CAA	A10951G	0.0033
Gln306	CAA	-	-

It is worth mentioning that most of the knowledge presented comes from previous studies using SARS-CoV M-pro but, given the high identity and structural similarity between it and the homologous in SARS-CoV-2, most of the results may be extrapolated. Only residues mediating known key interactions and/or those about which there are experimental results reported will be discussed in the following lines.

Ser1 from one protomer is involved in binding site shaping of the other protomer primarily via a well-known saline bridge with its N-terminal amino group to Glu166 and an intermolecular H-bond to Phe140, both from the other protomer (Figure 12) [12,45]. Here, we report these three residues as missense mutation-resistant residues. Glu166 role has been reported to be crucial because it mediates an important connection of the binding site with the dimer interface [45]. In fact, mutations of Glu166 in a R298A mutant blocks substrate-induced dimerization of M-pro *in vitro* and results in a complete loss of enzymatic activity [46]. Moreover, correct positioning of Phe140 plays a pivotal role in the hydrophobic

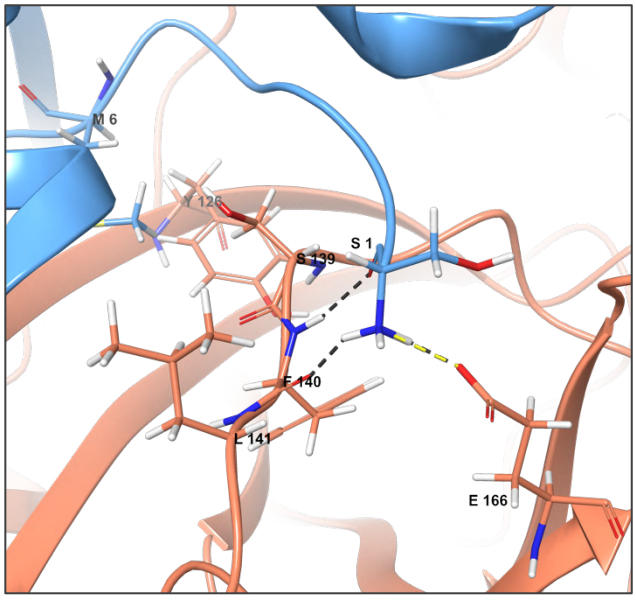

Figure 12 | Non-covalent interactions of the N-terminal Ser1 of one protomer and the other protomer, involving residues Phe140 and Glu166. One protomer is colored in blue and the other one is colored in orange. H-bond and salt bridges are represented as discontinuous lines in gray and yellow, respectively. Note that other important residues within this region are also depicted. This figure has been generated using Schrödinger Maestro and the structure of SARS-CoV-2 apo protein with PDBid code 6WQF.

pocket formed with Tyr126 [47]. This is presumably due to a major structural rearrangement of segment formed by residues 139-141, disorganization of which induces changes in the binding site and affects catalytic efficiency. Of note, residues 139-141 are part of the oxyanion loop and the former two are defined here as mutation coldspots. In fact, as proposed by Shi *et al.* [48] residues of the binding site which are also involved in dimerization might be the link between dimerization and catalysis. Thus, mutations of these residues make the binding site become more collapsed so that enzyme is unable to perform catalysis. Tyr126 also interacts hydrophobically with Met6, which is one of the residues on which more different mutations have been reported within the N-finger. Perhaps the most prominent mutation, M6L, is favoring stronger hydrophobic interactions within the hydrophobic pocket formed with Tyr126 and Phe140 and thereby is evolutionally selected. In addition, Hu *et al.* [49] reported that P140A mutation led to a dimeric conformation of M-pro but bearing a totally collapsed binding site.

A well-known interaction of the dimer interface involves side chains of Arg4 and Glu290, each one from a different protomer, which form a salt bridge (Figure 13). Moreover, truncation of residues 1-4 of the N-terminus of SARS-CoV-2 M-pro resulted in inability to dimerize and poor catalytic activity. Interestingly, truncation of only the first three residues had only a limited effect on the protease with 76% of wild-type enzyme activity, reinforcing the idea that Arg4 is indeed an important residue within the N-terminal region [50]. Chou *et al.* [51] reported that changes in pH and salt concentration led to a decrease in M-pro dimerization and catalytic efficiency. It is worth mentioning that Arg4 is not found among found coldspots of M-pro, although the only missense mutation found is to a Lys residue, which might not have an important effect in the interaction since a salt bridge is still possible. In fact, the substitution to a lysine could induce a tighter packing of the M-pro dimer due to the smaller residue. Nonetheless, it is also true that mutations of Arg4 might be more tolerated than mutations of Glu290, because mutation to Ala of the former had 5-fold decrease in dimerization of M-pro and a moderate effect on catalytic efficiency whereas mutations of the latter had dramatic effects on both aspects [51]. Besides Glu290, Arg4 has been found to interact with Lys137 –which is reported here as mutation coldspot– of the other protomer forming a H-bond (Figure13) [52].

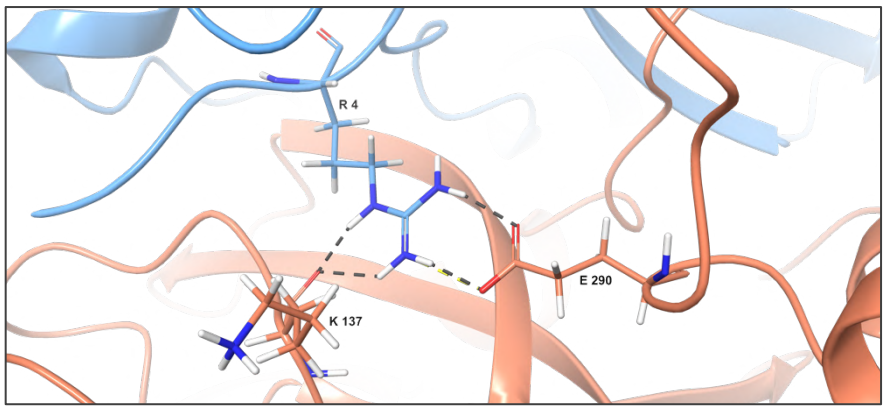

Figure 13 | Non-covalent interaction network between Arg4 from one protomer and Glu290 and Lys 137 from the othe protomer. One protomer is colored in blue and the other one is colored in orange. H-bond and salt bridges are represented as discontinuous lines in gray and yellow, respectively. This figure has been generated using Schrödinger Maestro and the structure of SARS-CoV-2 apo protein with PDBid code 6WQF.

Gly11 from one protomer interacts with the carboxyl group of Glu14 of the other protomers, via an intermolecular H-bond (Figure 14) [53]. Both residues are reported as mutation coldspots. Also, this interaction appears to be symmetric: Gly11 from protomer A interacts with Glu14 from protomer B and vice versa. Interestingly, mutation of Gly11 to Ala results in a dramatic change in the helical conformation of the first helix in domain I that abolishes the ability of the N-finger to squeeze in its usual position. This mutation brings about a tragic decrease to <1% the wild-type enzymatic activity [45]. Gly11, which is part of a short helical segment spanning residues 11-16, is indeed important in dimer formation as it mediates several interactions with the same helical segment of the other protomer. For instance, Ser10, another mutation coldspot, is known to mediate two H-bonds (one with the -NH of the main chain and the other with its side chain hydroxyl group) with the homologous serine in the other protomer. Besides, Cys16, which contains a protonated thiol group on its side chain, is also conservated and it is known to mediate important interactions with Ser10, Gly11 and Glu14 in SARS-CoV-2 dimer interface [52]. In fact, there are at least 13 intermolecular H-bonds between the N-terminus of one protomer and the same region other, which might explain the high level of conservation observed [14].

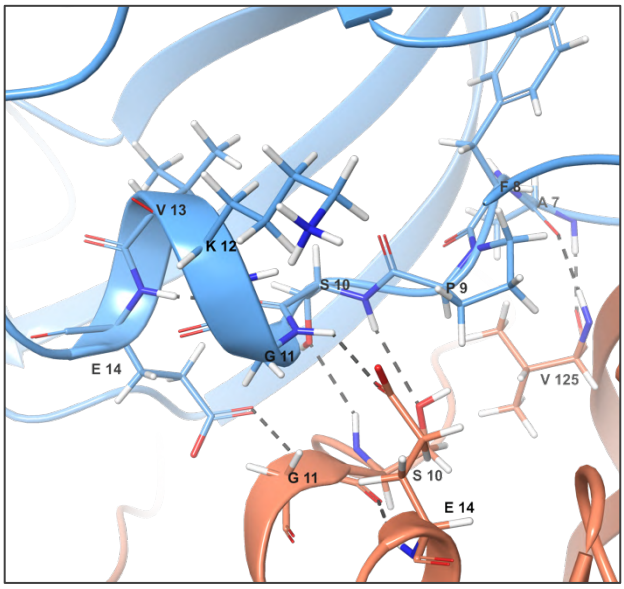

Figure 14 | Non-covalent interactions between Gly11 and Glu14 and neighboring residues. One protomer is colored in blue and the other one is colored in orange. H-bonds are represented as discontinuous lines in gray. This figure has been generated using Schrödinger Maestro and the structure of SARS-CoV-2 apo protein with PDBid code 6WQF.

Bacha *et al.* [54] identified a cluster of three conserved serine residues in some M-pro of different CoVs: Ser139, Ser144 and Ser147 (Figure 15, see also Figure 16). These residues are important in binding site shaping and orientation of the ligand. In this analysis, we found that Ser139 and Ser147 can be considered as mutation coldspots. Despite some mutations have been found in position 144 (*i.e.,* two different mutations that result in a substitution for a Glu residue and a third one to a Lys), these are found at very low frequencies (*i.e.,* found 2, 2 and 1 times, respectively). Moreover, in another study [55] they showed that mutation in each of these three residues had a notable effect on enzymatic activity. S139A substitution still had some activity, in line with what was also reported by Hu *et al.* [49]. Specifically, it seems that although impairing dimerization, some dimer structures can still be found, which may account for the observed activity. Surprisingly, S147A had the most dramatic effect and decreased about 150-fold the enzymatic activity of SARS-CoV M-pro. Ser147 is a residue which is buried between domains I and II just behind the binding pocket and it is located far away (9Å) from the dimer interface. However, S147A mutation resulted in the inability to dimerize of M-pro, suggesting

a long-cooperative interaction network between the dimer interface and the binding site. Ser147 might channel important interactions within this network. For instance, Ser147 might interact with His163 and the side chain of Ser144. Also, Asn28, another reported mutation coldspot, mediates key interactions within the M-pro structure. In fact, mutation of this residue present in the vicinity of these residues (*e.g.,* Ser 147, Ser144) results in a complete loss of the enzymatic activity –probably due to its role in correct positioning of Cys145– and a dramatic decrease (*i.e.,* roughly 20-fold) in the dimerization dissociation constant [47].

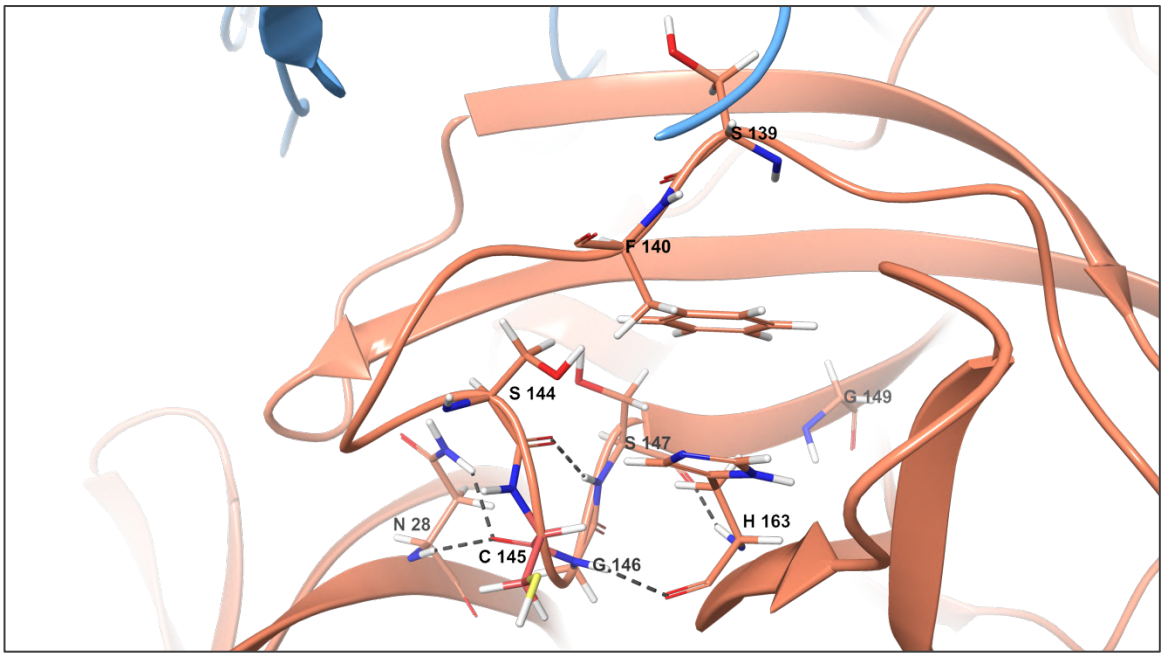

Figure 15 | Conserved serine cluster (Ser139, Ser 144 and Ser147) non-covalent interactions and neighboring coldspots. H-bonds are represented as discontinuous lines in gray. Note that Cys145, which is not a mutation coldspot, is also depicted in reddish orange, as it interacts with Asn28. This figure has been generated using Schrödinger Maestro and the structure of SARS-CoV-2 apo protein with PDBid code 6WQF.

Gly146 and His163 are two of the few residues of the binding site which can be considered as coldspots. In fact, Gly146 and His163 are probably interacting through an H-bond (Figure 15). While the role of Gly146 is still unclear, His163 has a prominent role in ligand binding [56]. For Gly146, we can hypothesize that due to its localization at the end of the oxyanion loop in a very restricted space, mutation to other residues might disturb the conformation of the oxyanion hole and therefore inhibit enzyme activity. Nonetheless, as said, this is only a hypothesis. For His163, *in silico* mutagenesis of this residue showed reduced affinity for ligands in MD simulations [56]. Two other residues situated in the vicinity of the binding site are also mutational coldspots: Cys44, which encases the active site along with the flexible loop described by Bzówka *et al.* [42], and Phe150 which is indeed important for binding site shaping [44]. Pro39 and Leu 27 also mediate interdomain nonpolar interactions that, along with other non-conserved residues, define the hydrophobic S2 pocket [15].

Finally, the C-terminal domain of SARS-CoV-2 M-pro is greatly conserved. Considering the important role that the last helix spanning residues 293-306 has in dimerization (discussed above and reviewed in detail in [57]) it is unsurprising that most residues are mutation coldspots. For instance, Shi and Song

[58] showed that mutations in residues Asp289 and Gln299 were sufficient to disrupt SARS-CoV M-pro dimerization. This agrees with a posterior report from Lin *et al.* [59] in which mutation of Gln299 (but also Arg298, which is not a mutation coldspot because it has three missense mutations reported) increased up to 4,000-fold the dissociation constant of the wild-type. Interestingly, mutation of Arg298 to Lys, which is the most frequently found mutation on this position in SARS-CoV-2, had no significant changes in k_{cat} compared to wild-type M-pro. In addition, the last residue (Gln306) is conserved in SARS-CoV-2, probably not only for its interactions, but also and mainly, because of the requirement of the S1 subsite of M-pro, which makes it able to process itself.

4.3.2. Conservation of mutation coldspots in other CoVs

To further investigate the biological importance of M-pro mutation coldspots, it could also be interesting to assess the conservation of SARS-CoV-2 M-pro among different CoVs (Figure 16). To this end, a multiple sequence alignment was performed. It is indeed true that M-pro is a highly conserved protein among different CoVs, with conserved clusters in between more variable regions. In fact, it seems clear that both N- and C- termini are conserved, with some residues, belonging to the binding site, which are also fully conserved. Both the catalytic His41 and Cys145 are also conserved.

Virtually all mutation coldspots (52 out of 54) were conserved in at least 50% of the considered sequences, highlighting again the importance of these residues. Moreover, 31 out of 54 (57.4 %) are totally conserved among all CoVs. However, it is worth highlighting that, as occurs with residue conservation in general, mutation coldspots are clustered in both ends of the proteins. Interestingly, Ser62, which is indeed conserved in all SARS-CoV-2 genomes, is only found in SARS-CoV M-pro apart from SARS-CoV-2. Due to its situation in the M-pro structure in the outer and more distal part of domain I, a clear function cannot be inferred. Although it might be possible that such conservation is just due to probability, a fully conserved residue after one year of evolution should have an important function, as it occurs with most of coldspots (discussed in detail in section 4.3.1). Also, among mutation-resistant residues conserved in almost all CoVs, a conserved GSCGSxG motif is present (residues 143-149 of the reference SARS-CoV-2 M-pro sequence), which was identified as important for initiating catalysis in SARS-CoV and MERS-CoV [60]. Also, it was reported that residues in positions 2, 4, 286 and 295, were known to have important roles at the dimer structure of PEDV, TGEV and hCoV-229E [44]. Of special interest is Leu286 from SARS-CoV-2. Recalling from the introduction, Thr285Ala and Ile286Leu are two of the 12 substitutions found between SARS-CoV and SARS-CoV-2. Both residues are highly variable among the aligned sequences, suggesting a systematical mutation pattern of this residue. The fact that SARS-CoV-2 Leu286 is still conserved supports that those residues were specifically and evolutionally selected to induce a tighter packing of the M-pro dimer.

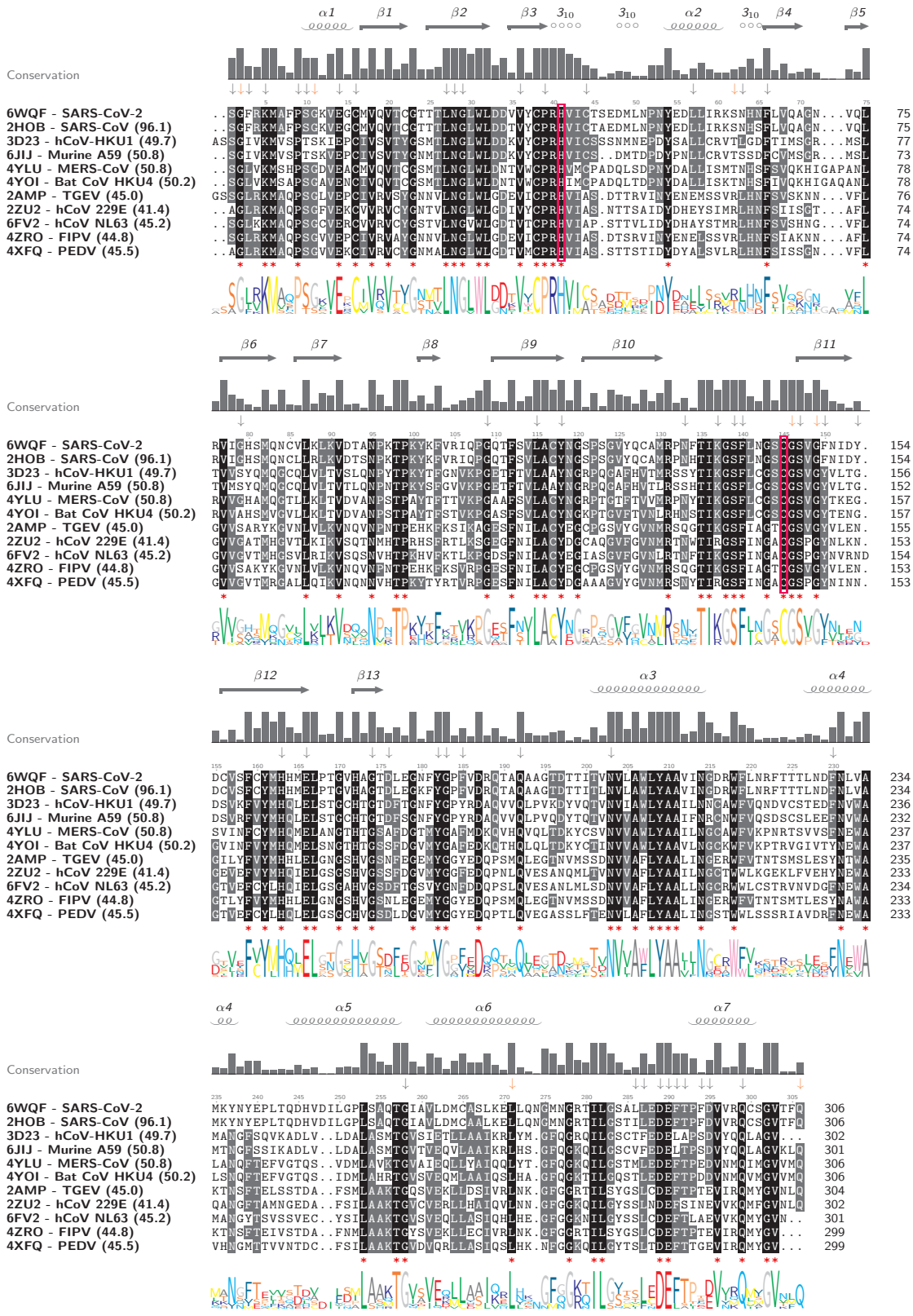

Figure 16 | Multiple sequence alignment (MSA) of SARS-CoV-2 M-pro and other CoVs. Name of different CoVs, an associated PDB structure and the identity percentage calculated by pairwise alignment to each sequence with SARS-CoV-2 M-pro (in brackets) are shown as title of each sequence. The secondary structure assigned to SARS-CoV-2 M-pro (PDBid 6WQF) by STRIDE is displayed at the upper part of the MSA, along with the degree of conservation represented by a bar plot. Mutation coldspots are highlighted using down-facing arrows and the distinction between missense mutation-resistant (gray) and conserved residues (orange) is made. Totally conserved residues are highlighted in black, whereas residues conserved in at least 50% of sequences are highlighted in gray. Less conserved residues are not highlighted. A red asterisk also indicates totally conserved residues. A sequence logo is displayed at the bottom part and residues are colored according to RasMol color scheme. Abbreviations: α, alpha helix; β, beta strand; 3_{10}, 3_{10} helix; hCoV, human CoV; TGEV, Transmissible gastroenteritis virus; FIPV, Feline infectious peritonitis virus; PEDV, Porcine epidemic diarrhea virus.

5. CONCLUSION

The present study provides a comprehensive view of the mutational profile of SARS-CoV-2. Mutational profiling enables the recognition of important features of a genome, as deleterious mutations are not evolutionally conserved, whereas mutations favoring viral fitness are indeed selected. For instance, the present study shows that the ribosomal frameshifting signal, on which SARS-CoV-2 depends to synthetize proteins at a regulated level, is highly conserved. Also, at a genomic level, the observed variation so far is clearly located primarily in both ends of the genome and the most conserved regions fall in the center of the genome, where proteins that are crucial for viral replication are encoded.

SARS-CoV-2 M-pro plays a pivotal role in viral replication, reason why it is considered as a main target to inhibit its replication. Therefore, gaining knowledge about possible targets within M-pro is of special interest. Here, a detailed analysis of the mutations found in M-pro is presented, specially focused on the biological importance of mutation-resistant residues or mutation coldspots. Moreover, an extensive review of reported and putative roles of the 54 herein found mutation coldspots suggests that specifically target some of these residues might allow a potent inhibition of viral replication which would not lose effectiveness over time as well. Also, it was found that mutation coldspots were mainly mediating interprotomer interactions or funneling interaction networks from the binding site towards the dimerization surface and vice versa. Interestingly, the binding site of M-pro tolerates mutations quite well. Therefore, this analysis proposes targeting M-pro dimerization or selected residues rather than the binding site, as occurs in classical *in silico* approaches, as an alternative and promising strategy for M-pro inhibition.

In summary, the present work paves the way for further studies regarding SARS-CoV-2 genomic analyses. The present data also permits the same analysis conducted with M-pro to be conducted with other key proteins. Also, with the M-pro analysis in mind, a new paradigm regarding its inhibition is presented. Gaining insight into the important residues of M-pro is valuable in other fields such as target-directed drug design. Plus, effects of mutations in coldspots, at least *in silico* by means of MD simulations, could be performed to support the theoretical suggestions presented here and it is indeed in the natural steps of this work to be continued.

6. ACKNOWLEDGEMENTS

I would like to first thank my tutor Dr. Gerard Pujadas. Also, my thanks and appreciations go to my other supervisor, Dr. Santi Garcia-Vallvé. They gave me the opportunity to enter to the fascinating world of bioinformatics and they have been always there happy to help me. I do really appreciate the trust they have placed on me since the first day as well as their commitment. Also, I would like to express my gratitude to all the members of the Cheminformatics and Nutrition research group, it has been a pleasure to be part of this team.

7. REFERENCE LIST

[1] Hu B, Guo H, Zhou P, Shi Z-L. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol 2021;19:141–54.

[2] Chan JF-W, Kok K-H, Zhu Z, Chu H, To KK-W, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect 2020;9:221–36.

[3] Zhao J, Cui W, Tian BP. The Potential Intermediate Hosts for SARS-CoV-2. Front Microbiol 2020;11:580137.

[4] Zhou P, Yang X Lou, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579:270–3.

[5] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020;395:565–74.

[6] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020;579:265–9.

[7] Omar SI, Zhao M, Sekar RV, Moghadam SA, Tuszynski JA, Woodside MT. Modeling the structure of the frameshift-stimulatory pseudoknot in SARS-CoV-2 reveals multiple possible conformers. PLOS Comput Biol 2021;17:e1008603.

[8] Prates ET, Garvin MR, Pavicic M, Jones P, Shah M, Demerdash O, et al. Potential Pathogenicity Determinants Identified from Structural Proteomics of SARS-CoV and SARS-CoV-2. Mol Biol Evol 2021;38:702–15.

[9] Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. Nature 2020;582:289–93.

[10] Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved a-ketoamide inhibitors. Science 2020;368:409–12.

[11] Huang Y, Yang C, Xu X feng, Xu W, Liu S wen. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin 2020;41:1141–9.

[12] Cannalire R, Cerchia C, Beccari AR, Di Leva FS, Summa V. Targeting SARS-CoV-2 Proteases and Polymerase for COVID-19 Treatment: State of the Art and Future Opportunities. J Med Chem 2020.

[13] Gimeno A, Mestres-Truyol J, Ojeda-Montes MJ, Macip G, Saldivar-Espinoza B, Cereto-Massagué A, et al. Prediction of novel inhibitors of the main protease (M-pro) of SARS-CoV-2 through consensus docking and drug reposition. Int J Mol Sci 2020;21(11):3793.

[14] Kneller DW, Phillips G, O HM, Jedrzejczak R, Stols L, Langan P, et al. Structural plasticity of SARS-CoV-2 3CL M pro active site cavity revealed by room temperature X-ray crystallography. Nat Commun 2020; 11: 3202

[15] Suárez D, Díaz N. SARS-CoV-2 Main Protease: A Molecular Dynamics Study. J Chem Inf Model 2020;60:5815–31.

[16] Xue X, Yu H, Yang H, Xue F, Wu Z, Shen W, et al. Structures of Two Coronavirus Main Proteases: Implications for Substrate Binding and Antiviral Drug Design. J Virol 2008;82:2515–27.

[17] Ramos-Guzmán CA, Ruiz-Pernía JJ, Tuñón I. Unraveling the SARS-CoV-2 Main Protease Mechanism Using Multiscale Methods. ACS Catal 2020;10:12544–54.

[18] Chang GG. Quaternary structure of the SARS coronavirus main protease. Molecular Biololgy of the SARS-Coronavirus, Springer Berlin Heidelberg; 2010, p.115–28.

[19] Gahlawat A, Kumar N, Kumar R, Sandhu H, Singh IP, Singh S, et al. Structure-Based Virtual Screening to Discover Potential Lead Molecules for the SARS-CoV-2 Main Protease. J Chem Inf Model 2020;60:5781–93.

[20] Muramatsu T, Kim YT, Nishii W, Terada T, Shirouzu M, Yokoyama S. Autoprocessing mechanism of severe acute respiratory syndrome coronavirus 3C-like protease (SARS-CoV 3CLpro) from its polyproteins. FEBS J 2013;280:2002–13.

[21] Team R Development Core. R: A Language and Environment for Statistical Computing 2021.

[22] Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016.

[23] Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. Msa: An R package for multiple sequence alignment. Bioinformatics 2015;31:3997–9.

[24] Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 2004;32:W500–2.

[25] Carrasco-Hernandez R, Jácome R, Vidal YL, de León SP. Are RNA viruses candidate agents for the next global pandemic? A review. ILAR J 2017;58:343–58.

[26] Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: Patterns and determinants. Nat Rev Genet 2008;9:267–76.

[27] Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, et al. Infidelity of SARS-CoV Nsp14-Exonuclease Mutant Virus Replication Is Revealed by Complete Genome Sequencing. PLoS Pathog 2010;6:e1000896.

[28] Wang R, Hozumi Y, Zheng YH, Yin C, Wei GW. Host immune response driving SARS-CoV-2 evolution. Viruses 2020;12:1–20.

[29] Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med 2020;18:1–9.

[30] Chand GB, Banerjee A, Azad GK. Identification of novel mutations in RNA-dependent RNA polymerases of SARS-CoV-2 and their implications on its protein structure. PeerJ 2020;8:e9492.

[31] Zhang L, Jackson CB, Mou H, Ojha A, Peng H, Quinlan BD, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun 2020;11:1–9.

[32] Isabel S, Graña-Miraglia L, Gutierrez JM, Bundalovic-Torma C, Groves HE, Isabel MR, et al. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. Sci Rep 2020;10:1–9.

[33] Yurkovetskiy L, Wang X, Pascal KE, Tomkins-Tinch C, Nyalile TP, Wang Y, et al. Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. Cell 2020;183:739-51.

[34] Kelly JA, Olson AN, Neupane K, Munshi S, Emeterio JS, Pollack L, et al. Structural and functional conservation of the programmed −1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). J Biol Chem 2020;295:10741–8.

[35] Jaroszewski L, Iyer M, Alisoltani A, Sedova M, Godzik A. The interplay of SARS-CoV-2 evolution and constraints imposed by the structure and functionality of its proteins. Preprint. BioRxiv 2020;2020.08.10.244756.

[36] Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. NextStrain: Real-time tracking of pathogen evolution. Bioinformatics 2018;34:4121–3.

[37] Liu S, Shen J, Fang S, Li K, Liu J, Yang L, et al. Genetic Spectrum and Distinct Evolution Patterns of SARS-CoV-2. Front Microbiol 2020;11:1–14.

[38] Parvez MSA, Rahman MM, Morshed MN, Rahman D, Anwar S, Hosen MJ. Genetic analysis of SARS-CoV-2 isolates collected from Bangladesh: Insights into the origin, mutational spectrum and possible pathomechanism. Comput Biol Chem 2021;90:107413.

[39] Chan KF, Koukouravas S, Yeo JY, Koh DWS, Gan SKE. Probability of change in life: Amino acid changes in single nucleotide substitutions. BioSystems 2020;193–194:104135.

[40] Martin RW, Butts CT, Cross TJ, Takahashi GR, Diessner EM, Crosby MG, et al. Sequence characterization and molecular modeling of clinically relevant variants of the SARS-CoV-2 main protease. Biochemistry 2020;59:3741–56.

[41] Amamuddy OS, Verkhivker GM, Bishop ÖT. Impact of early pandemic stage mutations on molecular dynamics of SARS-CoV-2 MPro. J Chem Inf Model 2020;60:5080–102.

[42] Bzówka M, Mitusińska K, Raczyńska A, Samol A, Tuszyński JA, Góra A. Structural and evolutionary analysis indicate that the sars-COV-2 mpro is a challenging target for small-molecule inhibitor design. Int J Mol Sci 2020;21(9):3099.

[43] Badua CLDC, Baldo KAT, Medina PMB. Genomic and proteomic mutation landscapes of SARS-CoV-2. J Med Virol 2021;93:1702–21.

[44] Krishnamoorthy N, Fakhro K. Identification of mutation resistance coldspots for targeting the SARS-CoV2 main protease. IUBMB Life 2021;73:670–5.

[45] Chen S, Hu T, Zhang J, Chen J, Chen K, Ding J, et al. Mutation of Gly-11 on the dimer interface results in the complete crystallographic dimer dissociation of severe acute respiratory syndrome coronavirus 3C-like protease: Crystal structure with molecular dynamics simulations. J Biol Chem 2008;283:554–64.

[46] Cheng SC, Chang GG, Chou CY. Mutation of glu-166 blocks the substrate-induced dimerization of SARS coronavirus main protease. Biophys J 2010;98:1327–36.

[47] Barrila J, Gabelli SB, Bacha U, Amzel LM, Freire E. Mutation of Asn28 Disrupts the Dimerization and Enzymatic Activity of SARS 3CL pro ,. Biochemistry 2010;49:4308–17.

[48] Shi J, Sivaraman J, Song J. Mechanism for Controlling the Dimer-Monomer Switch and Coupling Dimerization to Catalysis of the Severe Acute Respiratory Syndrome Coronavirus 3C-Like Protease. J Virol 2008;82:4620–9.

[49] Hu T, Zhang Y, Li L, Wang K, Chen S, Chen J, et al. Two adjacent mutations on the dimer interface of SARS coronavirus 3C-like protease cause different conformational changes in crystal structure. Virology 2009;388:324–34.

[50] Hsu WC, Chang HC, Chou CY, Tsai PJ, Lin PI, Chang GG. Critical assessment of important regions in the subunit association and catalytic action of the severe acute respiratory syndrome coronavirus main protease. J Biol Chem 2005;280:22741–8.

[51] Chou C-Y, Chang H-C, Hsu W-C, Lin T-Z, Lin C-H, Chang G-G. Quaternary Structure of the Severe Acute Respiratory Syndrome (SARS) Coronavirus Main Protease. Biochemistry 2004;43:14958–70.

[52] Kneller DW, Phillips G, Weiss KL, Pant S, Zhang Q, O'Neill HM, et al. Unusual zwitterionic catalytic site of SARS-CoV-2 main protease revealed by neutron crystallography. J Biol Chem 2020;295:17365–73.

[53] Xia B, Kang X. Activation and maturation of SARS-CoV main protease. Protein Cell 2011;2:282–90.

[54] Bacha U, Barrila J, Velazquez-Campoy A, Leavitt SA, Freire E. Identification of Novel Inhibitors of the SARS Coronavirus Main Protease 3CLpro. Biochemistry 2004;43:4906–12.

[55] Barrila J, Bacha U, Freire E. Long-Range Cooperative Interactions Modulate Dimerization in SARS 3CL pro. Biochemistry 2006;45:14908–16.

[56] Weng YL, Naik SR, Dingelstad N, Lugo MR, Kalyaanamoorthy S, Ganesan A. Molecular dynamics and in silico mutagenesis on the reversible inhibitor-bound SARS-CoV-2 main protease complexes reveal the role of lateral pocket in enhancing the ligand affinity. Sci Rep 2021;11:1–22.

[57] Goyal B, Goyal D. Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. ACS Comb Sci 2020;22:297–305.

[58] Shi J, Song J. The catalysis of the SARS 3C-like protease is under extensive regulation by its extra domain. FEBS J 2006;273:1035–45.

[59] Lin PY, Chou CY, Chang HC, Hsu WC, Chang GG. Correlation between dissociation and catalysis of SARS-CoV main protease. Arch Biochem Biophys 2008;472:34–42.

[60] Wang H, He S, Deng W, Zhang Y, Li G, Sun J, et al. Comprehensive Insights into the Catalytic Mechanism of Middle East Respiratory Syndrome 3C-Like Protease and Severe Acute Respiratory Syndrome 3C-Like Protease. ACS Catal 2020;10:5871–90.

8. SELF-ASSESSMENT

The first time I found myself in front of a computer with tons of data to work with, I thought it would be impossible for me to eventually reach my objective. But I faced this as an enormous challenge. It is worth bearing in mind that I have been working in the QiN group for around two years, but before that bioinformatics were totally new to me. In fact, I had to learn programming from zero.

I honestly consider my experience in the QiN group, in general, and with the present work, in specific, as indeed valuable both at a professional and personal level. I have learnt to work autonomously but also be part of a team and share my results with my colleagues. Also, I do appreciate the way that my supervisors let me organize my work, make suggestions and contributions, and bring about the final result I present here. Plus, I learnt a lot about method, critical thinking and the use of scientific language. Not to mention all the professional abilities I have acquired while performing this work: programming, managing data, interpreting and presenting results... Another think I consider to be crucial is learning from your errors and being capable of admit them. This work has been a real trial and error race, in which starting from scratch was not an uncommon thing. Nonetheless, I found this experience more important than all the success one could think of.

After all, growth is the word that defines this work and my stay in the QiN research group. Every second invested in here has paid off.