U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

2.º
CICLO

FCUP
2021

U.PORTO

A lesson from Coronaviridae: Predictive models for SARS-CoV-2 infection evolution and identification of potential future threats

Ricardo Jorge Vieira Soares

FC

A lesson from *Coronaviridae*: Predictive models for SARS-CoV-2 infection evolution and identification of potential future threats

Ricardo Jorge Vieira Soares

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Biologia Celular e Molecular

2021

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

A lesson from *Coronaviridae*: Predictive models for SARS-CoV-2 infection evolution and identification of potential future threats

Ricardo Jorge Vieira Soares
Cell and Molecular Biology Master's Degree
Biology Department, FCUP
2021

Supervisor

Jorge Manuel De Sousa Basto Vieira
Main researcher at IBMC

Co-supervisor

Cristina Alexandra Gonçalves Paula Vieira,
Auxiliary researcher at IBMC

Institute Address

IBMC – Instituto de Biologia Molecular e Celular
I3s building – Rua Alfredo Allen nº 208 4200-135 Porto, Portugal



Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

“For small creatures such as we the vastness is bearable only through love.”

For this and so much more, I dedicate this dissertation to my Maria. 

Agradecimentos

Contrastando com o restante texto, onde meramente descrevo o que observei e deduzi, aqui apenas poderei colocar um fragmento daquilo que sinto sobre quem me faz companhia nesta jornada.

Agradeço ao meu orientador, Professor Doutor Jorge Vieira por partilhar comigo as frustrações de quando surgia um obstáculo, por me ter feito sentir o seu entusiasmo nas vitórias e por ter sempre um olhar otimista. Agradeço à minha coorientadora Professora Doutora Cristina Vieira pelo seu olhar pragmático e pela sua exigência pelo meu trabalho. Obrigado a ambos por terem a coragem e vontade em apostar e confiar um tema novo a um aluno ainda mais novo. Obrigado por todas as conversas e discussões, onde sempre pude aprender. Um muito obrigado por me terem feito baixar a guarda ao ponto de não escrever as unidades num exame, aí soube que o meu caminho teria que passar por aqui.

Um obrigado, claro está, a quem se fez importante na minha vida, fazendo tanto parte de mim como eu deles. Obrigado Catarina por durante 5 anos me mostrares a tua força e resiliência. Obrigado Diogo por me deixares ver a pessoa sensível e bonita que és. Obrigado Catarina F. por não deixares cair planos loucos, os insetos são o caminho. Obrigado André pela companhia e por nunca hesitares em ajudar. Obrigado Jéssica por descomplicares o mestrado e por seres tão perturbada como eu a trabalhar. Obrigado Joana por estimulares o meu humor negro. Obrigado Rute por reparares sempre que a minha roupa combina com os óculos. Obrigado Rita por te despedires dizendo “diverte-te”. À Colete e à Sofia, obrigado pela integração e pelas saídas e jantaradas.

A toda a minha família, pais, avós, tios e prima, obrigado por me terem dado tudo. Obrigado por todo o vosso amor. Só assim sou o que sou, como sou. Obrigado por me terem dado liberdade para que possa traçar este meu caminho. Que este seja apenas mais um passo para me tornar num motivo do vosso orgulho.

Ao Gonçalo, que marcou o primeiro ponto de viragem da minha vida, agradeço tudo. Desde o início que foste o irmão que nunca tive e o meu ponto de inspiração.

À minha Maria, dedico este estudo e este curso. Entraste na pior fase e tornaste-a na melhor, deste-me um novo rumo, fizeste-me voltar a acreditar em mim e a ver a beleza do mundo. Contigo a meu lado, não há nada que não consigamos fazer. Que seja o futuro o lugar onde seremos felizes, para sempre.

A todos, isto é vosso!

*Agradeço ao SARS-CoV-2, pois, sem ele não seria possível realizar este estudo.

Resumo

A família *Coronaviridae* é constituída por espécies capazes de infetar um largo espetro de hospedeiros. Em menos de 20 anos, três destas espécies foram transmitidas a humanos, provocando doença e morte. A mais recente, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), deu origem a uma pandemia em 2020. A sua grande distribuição e prevalência na população permitiu, em 2021, o aparecimento de novas variantes com diferentes taxas de transmissibilidade. Além disso, este vírus mostrou não ter um hospedeiro único, com casos documentados de infeção em diferentes espécies de animais.

Este estudo aplicou uma abordagem bioinformática exaustiva sobre a problemática *Coronaviridae* que revelou padrões que podem ser utilizados para obter uma larga variedade de previsões, sobre transmissão de SARS-CoV-2 entre espécies, a taxa de transmissibilidade de variantes de SARS-CoV-2 e a identificação de variantes positivamente selecionadas. Acerca da transmissão de SARS-CoV-2 entre espécies, o método aqui desenvolvido, baseado nas observações aqui descritas de que existe uma diferença nos resíduos específicos da ACE2 que estabelecem interface com a proteína S entre espécies suscetíveis e não suscetíveis a SARS-CoV-2, identificou uma lista de espécies candidatas suscetíveis à infeção por SARS-CoV-2 que é provavelmente mais próxima da realidade do que aquela já descrita na literatura. Ademais, espécies que utilizam a ACE2 apresentam um TM-score do seu RBM estatisticamente mais elevado, quando emparelhado com o RBM de SARS-CoV-2, do que aquelas incapazes de a usar, tornando-o um método útil para determinação do uso da ACE2 por CoVs. Adicionalmente, aqui é mostrado que, em humanos, a taxa de transmissibilidade de variantes de SARS-CoV-2 está correlacionada com o número de resíduos em interface da ACE2. Assim sendo, um protocolo simples e rápido foi desenvolvido para prever a transmissibilidade de variantes de SARS-CoV-2. Finalmente, a análise de locais positivamente selecionados em bases de dados de SARS-CoV-2 revelou que estes correspondem em grande parte a variantes de interesse e preocupação.

Abstract

The Coronaviridae family is composed of species able to infect a broad range of hosts. In less than 20 years, three of these species have been transmitted to humans, causing disease and death. The most recent, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), turned into a pandemic in 2020. The large distribution and prevalence in the population allowed the birth of new variants with different transmissibility rates in 2021. Furthermore, this virus showed not to be host specific, with cases of different animal species infected being documented.

The exhaustive bioinformatics study here conducted on *Coronaviridae* revealed patterns that can be used to make a wide variety of useful predictions, regarding between species transmission of SARS-CoV-2, the transmissibility rate of SARS-CoV-2 variants, and the identification of positively selected variants. Regarding the between species transmission of SARS-CoV-2, the method here developed, based on the observation here made that there are differences in the specific ACE2 residues that establish the interface with S protein in SARS-CoV-2 susceptible and not susceptible species, identified a list of candidate species susceptible to SARS-CoV-2 infection that is likely closer to reality than the one already reported in the literature. Moreover, species that use ACE2 presented a statistically higher TM-score of their RBM when paired with the RBM of SARS-CoV-2 when compared with the ones that don't use it, making it also a useful method for determining the use of ACE2 by CoVs. Additionally, it is here shown that, in humans, the transmissibility rate of SARS-CoV-2 variants is correlated to the number of ACE2 interfacing residues. Therefore, a simple and fast protocol was developed to predict the transmissibility of SARS-CoV-2 variants. Finally, the positively selected amino acid sites identified in SARS-CoV-2 datasets, revealed that these correspond mostly to the variants of concern and of interest.

Key-words

SARS-CoV-2, *Coronaviridae*, spike glycoprotein, angiotensin converting enzyme 2 (ACE2), I-TASSER, HADDOCK, PDBePISA, IPSSA, prediction model, interfacing residues.

Index

Agradecimentos i
Resumo iii
Abstract iv
Key-words iv
Index v
Table Index vii
Figure Index viii
Abbreviation list x
1. Introduction 1
1.1 *Coronaviridae* 1
1.2 XXI century: from outbreaks to a Pandemic 2
1.3 Spillover: Origin and Intermediate Hosts 2
1.4 SARS-CoV-2 origin 3
1.5 Surface glycoprotein 4
1.5.1 S protein architecture 5
1.6 SARS-CoV-2 variants 6
1.7 Angiotensin Converting Enzyme 2 7
1.7.1 ACE and ACE2 8
1.8 Interface: ACE2 – SARS-CoV/2 S protein 9
1.9 Protein – Protein docking 10
1.9.1 Fully automated computer algorithms 10
1.9.2 Docking algorithms 10
1.10 Positively selected sites (PSS) 11
2. Objectives 12
3. Methods and materials 13
3.1 Sequence retrieval 13
3.1.1 ACE2 and reference S protein 13
3.1.2 *Coronaviridae* S proteins 14
3.1.3 SARS-CoV-2 variants' S protein 14
3.2 Phylogenetic approach 15
3.2.1 ACE2 gene tree 15
3.2.2 S protein gene tree 15
3.2.3 Consensus sequences 16
3.3 Protein-protein docking 16

3.3.1 Protein 3D structure prediction 16
3.3.2 HADDOCK 17
3.3.3 PDBePISA 18
3.4 Large Scale predictive model 18
3.5 S protein: Sites under positive selection 19
3.6 *Coronaviridae*'s use of ACE2 20
4. Results 20
4.1 ACE2 phylogeny 20
4.2 Consensus sequences 22
4.2.1 ACE2 22
4.2.2 SARS-CoV-2 S protein 22
4.3 SARS-CoV-2 variants 23
4.3.1 Number of interfacing residues 23
4.3.2 Hydrogen bonds and Salt bridges 24
4.3.3 Point mutations 25
4.4 SARS-CoV-2 variants and point mutation in core residues 28
4.5 SARS-CoV-2 predictive model for infection 29
4.5.1 SARS-CoV-2 – ACE2 core residues 31
4.5.2 Large scale predictive model 31
4.6 *Coronaviridae* 34
4.6.1 Identifying similar SARS-CoV-2's S protein in *Coronaviridae* 34
4.6.2 Use of ACE2 34
4.7 Positively selected sites 36
5. Discussion 37
5.1 Transmissibility rates of SARS-CoV-2 variants 37
5.2 ACE2 usage 38
5.3 Large scale susceptibility model 39
5.4 Positively selected sites 41
6. Conclusion 43
7. References 45
8. Supplementary data 60

Table Index

Table 1 – List of VOCs and VOIs identified by their Pango lineage identifier and their respective S proteins mutations. 15
Table 2 – List of core residues for both hACE2 and the S protein with frequencies equal or above 90% and 100%. For the S protein, the position number is based on the reference SARS-CoV-2 protein sequence. 28
Table 3 – List of core residues for both SARS-CoV-2 ‘s S protein and ACE2 of the positive control. The position number is based on the reference hACE2 protein sequence. 29
Table 4 – List of the intersected core residues ACE2 and S protein found for SARS-CoV-2 variants and SARS-CoV-2 susceptible species. 31
Table 5 – List of interfacing amino acids for the core residues in ACE2 and their respective common characteristic. Lack of a common characteristic signifies a very diverse site, with the exception of site 386, which is a conserved one. 32
Table 6 – Number of positively selected sites in SARS-CoV-2 and their respective frequency found using the codeML and FUBAR methods. 36
Table 7 – Positively selected sites of SARS-CoV-2 with frequencies above 25% using codeML and FUBAR methods. 37

Figure Index

Fig.1 – Neighbor-Joining ACE2 protein tree cladogram, with Anthozoa cluster used as root, collapsed by classes. Bootstraps values of 500 replicates are presented next to the branches, values under 80 are not shown. The yellow cluster represents the mammalian cluster with ACE2 sequences. The blue cluster depicts the mammalian cluster with ACE1 sequences. A total of 600 protein sequences were analysed in the final dataset conducted by MEGAX..... 21

Fig.2 – Cladogram of the bootstrap consensus sequence ACE2 amino acid tree by Neighbor-Joining method, collapsed by classes, with Hyperoartia as root. The bootstrap value of 500 replicates is presented, bootstrap values under 80 are not shown. Mammalian cluster is represented in orange. A total of 290 aa sequences were analysed conducted by MEGAX..... 22

Fig.3 – Number of interfacing residues of SARS-CoV-2 variants' S protein (white) and hACE2 (stripes), with their respective documented or expected transmissibility rate (dots). Results ordered by total number of interfacing residues. Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues. 23

Fig.4 – Correlation between the number of interfacing residues of SARS-CoV-2 variants and their expected and documented transmissibility rate. Image obtained using SPSS. 24

Fig.5 – Number of hydrogen bonds (black) and salt bridges (white) formed by the docking of hACE2 and the S protein of variants of concern and interest and with single mutations that have a documented or expected transmissibility rate. Results ordered by total number of interfacing residues. Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues. 25

Fig.6 – Effect of the point mutations N501Y (A) and D614G (B) to the ACE2 number of interfacing residues between hACE2 and S protein when applied to the reference sequence (grey), in tandem with variant defining groups of mutations (black) or added to existing variants (light grey). Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues. 26

Fig.8 – Effect of the point mutation A701V to the ACE2 number of interfacing residues between hACE2 and S protein when applied to the reference sequence (gray), in tandem with variant defining groups of mutations (black). Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues..... 27

Fig.9 – Effect of the point mutations to the ACE2 number of interfacing residues between hACE2 and S protein when applied to the reference sequence. Mutations in RBM are represented in black and mutations outside the RBM are represented in grey. Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues. 28
Fig.10 – ACE2 (A) and S protein (B) interfacing residues between SARS-CoV-2 S protein and Homo sapiens (white), susceptible species (black), non-susceptible species (grey), species with non-documented infection (stripes) and with hACE1 (squares). Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues. 30
Fig.11 – Mean number of ACE2 interfacing residues with SARS-CoV-2's S protein for susceptible (white) and non-susceptible species (grey). Error bars represent one standard deviation. Data processed in SPSS. 30
Fig.12 – Phylogenetic time tree of mammalian families adapted from Meredith et al. (2011). The number of candidates for SARS-CoV-2 – ACE2 recognition for a given family is displayed in a coloured square. The colour code correspond to the Damas et al. (2020) attributed score: blue – Very low; green – Low; yellow – Medium; orange – High; red – High. The number of species recognised for a given family with found candidates is also displayed (Myers et al., 2021). 33
Fig.13 – Average TM-score for the RBM of species of coronavirus that are able or unable to use the ACE2 protein to host cell entry, when paired with SARS-CoV-2. The score is normalized for the length of SARS-CoV-2's RBM. Error bars represent one standard deviation. Data processed in SPSS. 36

Abbreviation list

ACE1 – Angiotensin converting enzyme 1

ACE2 – Angiotensin converting enzyme 2

BSA – Buried surface area

CDC – Centers for Disease Control and Prevention

CDS – Coding sequence

CoV – Coronavirus

hACE2 – Human angiotensin converting enzyme 2

HADDOCK – High Ambiguity Driven protein-protein Docking

IPSSA – Integrated Positively Selected Sites Analyses

I-TASSER – Iterative Threading Assembly Refinement

MERS – Middle east respiratory syndrome coronavirus

NCBI – National Center for Biotechnology Information database

NMR – Nuclear magnetic resonance

ORF – Open reading frame

PD – Metallopeptidase domain

PDB – Protein data bank

PISA – Proteins, Interfaces, Structures and Assemblies

PSS – Positively selected sites

RBD – Receptor binding domain

RBM – Receptor binding motif

S protein – Surface glycoprotein

SARS – Severe acute respiratory syndrome

SEDA – Sequence Dataset builder

TMH – Single transmembrane helix

VHC – Variant of high consequence

ViPR – Virus pathogen resource

VOC – Variant of concern

VOI – Variant of interest

1. Introduction

1.1 *Coronaviridae*

Coronaviruses (CoVs) belong to the *Coronaviridae* family known for having large, 26-32 kb in length, single positive-sense RNA genomes. Members of this family typically show in the first two thirds at 5' the replicase gene, composed of ORF1A and 1b, followed by the spike, membrane and nucleocapsid encoding genes (de Groot et al., 2012). These zoonotic viruses, able to infect both animals and humans, were first identified in patients diagnosed with the common cold (Tyrrell & Bynoe, 1965). Belonging to the *Riboviria* realm, order *Nidovirales*, sub order *Cornidovirineae*, this family is divided in two subfamilies, five genera, 26 sub-genera and 46 species. Further, CoVs belong to the *Orthocoronavirinae* subfamily, that splits into four genera, those being α-, β-, δ- and γ-*coronavirus* (https://talk.ictvonline.org/taxonomy/ - accessed in January 02, 2021). The first two only infect mammals, with *α-coronavirus* having two of its 19 species able to infect humans, HCoV-229E and HCoV-NL63. The *β-coronavirus* has 14 species, among them HCoV-HKU1, HCoV-OC43, severe acute respiratory syndrome coronavirus (SARS-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV) and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), that are also able to infect humans (Gorbalenya et al., 2020).

HCoV-229E, HCoV-NL63, HCoV-HKU1 and HCoV-OC43 are known to be endemic and responsible for 15 to 30% of common colds, causing mild respiratory disease (Ahmad et al., 2020; Mesel-Lemoine et al., 2012). HCoV-229E and HCoV-OC43 are known since the 1960s, being first isolated in 1967 (Hamre & Procknow, 1966; McIntosh et al., 1967). As for HCoV-NL63, it was first isolated from a seven-month old infant in 2003 and shares 65% nucleotide identity with the other endemic human alphacoronavirus (van der Hoek et al., 2004). In 2005, the last of the four endemic HCoV, HCoV-OC43 was found in a 71-year-old man, showing the broad range of ages susceptible to infection (Woo et al., 2005). Species from the δ and γ-CoVs are not capable to infect humans, with the majority of its species targeting birds instead, however some are known to infect mammals (Woo et al., 2012). Interestingly, the type of illness *α-coronavirus* and *β-coronavirus* cause varies from respiratory in humans to gastroenteritis in other mammals, even though all endemic HCoVs have been detected in stool samples. Furthermore, a link between HCoV-OC43 and neurological diseases such as acute encephalomyelitis has been brought (Arbour et al., 1999; Morfopoulou et al., 2016; Murray et al., 1992).

1.2 XXI century: from outbreaks to a Pandemic

Recently, the world as faced three outbreaks caused by novel coronaviruses infecting humans. These, contrary to the endemic HCoVs, cause severe acute respiratory disease (Fehr et al., 2017). SARS-CoV was the first to emerge, in 2002 in southern China, and by the end of 2003 it had spread through 33 countries, accumulating 8096 confirmed infections, of which 774 (≈ 10%) resulted in death (da Silva et al., 2021; World Health Organization, 2015). About ten years later in 2012, MERS-CoV, causing a dangerous human disease, appeared in the Arabian Peninsula. As of September 2019, 27 countries have reported cases of Middle East respiratory syndrome (MERS) adding to a total of 2468 confirmed cases, of which 851 resulted in death. These numbers indicate a 34.4% global death rate, having Saudi Arabia as the prevalent location with more than 2000 cases and 771 deaths (World Health Organization, 2019). Lastly, in December 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged and by the 52nd week of 2020, 80 316 555 cases and 1 770 695 deaths have been reported, showing a death rate of 2.2% (European Centre for Disease Prevention and Control, 2020a). Its impact is not only a health issue, but also a humanitarian and economic one.

Due to the pandemic, the Global Economic Prospects predicted a 5.2% recession in global GDP in 2020 (World Bank, 2020). In Portugal, specifically, the expected GDP drop reaches the value of 8.0% coupling with an expected unemployment rate of 9.7 to 13.9% (Mamede et al., 2020).

1.3 Spillover: Origin and Intermediate Hosts

It is believed that all HCoV have originated in animals, even the endemic ones (Su et al., 2016). Bats are thought to be the origin of the endemic α-CoVs, as for the β-CoVs, their origin are rodents (Forni et al., 2017; Su et al., 2016). Nevertheless, coronaviruses found in *Vicugna pacos* and *Camelus dromedarius* are found to be close relatives to the HCoV-229E. Canids and artiodactyls, on the other hand, are found to be hosts for the sister clade to HCoV-OC43. This suggests that the viruses have a significant evolutionary history outside bats and rodents, prior to human cross species transmission (Corman et al., 2018; Cui et al., 2019).

As for SARS-CoV, shortly after the outbreak, it was believed that the raccoon dogs (*Nyctereutes procyonoides*) and Himalayan palm civets (*Paguma larvata*) were the main hosts involved in human transmission. These suspicions arose from the presence of genetically similar CoVs isolated from these species' samples present in the Guangdong province of China (Guan et al., 2003). Further studies reported the existence of a SARS-like CoV very similar to the human SARS-CoV isolated from *Rhinolophus affinis*, the Chinese horseshoe bat (BatCoV RaTG13) (Lau et al., 2005; Ye et al., 2020).

Additionally, this CoV was able to experimentally replicate in HeLa cells expressing virus receptor of multiple species. The existence of other Bat CoVs that, like the BatCoV RaTG13, share around 88 to 92% of nucleotide homology, points to bats as the main reservoir for SARS-CoV, with civets as intermediate hosts (Cui et al., 2019; Hu et al., 2017).

Likewise, bats are believed to be the evolutionary source of MERS-CoV, however the primary source is thought to be dromedary camels (Anthony et al., 2017; Guan et al., 2003). This hypothesis is supported by the ability of MERS-CoV to infect bat cells using the same method used in humans. Furthermore, strains isolated from camels are almost identical to those found in humans (Azhar et al., 2014; Sabir et al., 2016; Salata et al., 2020). The exposure from camels to this virus dates back to 1983, where serum samples revealed MERS infection, suggesting they may serve as a zoonotic reservoir since then (Cui et al., 2019; Müller et al., 2015). The presence of confirmed cases involving no interactions or contacts with camels caused doubts in the way that the virus transmits, whether between humans or by using other intermediate hosts (Samara & Abdoun, 2014; Ye et al., 2020). MERS-CoV was found to be phylogenetically related to *Tylonycteris* BatCoV-HKU4 and *Pipistrellus* BatCoV-HKU5, strengthening the hypothesis of its origin in bats (Lau et al., 2013). Being classified as a β-CoV lineage C, MERS-CoV has an identical genomic structure to the other members of this lineage, differing in its accessory proteins. However, using these variable proteins, no direct progenitor for MERS-CoV was found in bat species (Corman et al., 2014; Lau et al., 2013; Yang et al., 2014).

1.4 SARS-CoV-2 origin

Understanding how the species boundary is crossed will help to prevent future zoonotic events. In the case of pre-adaptation in animals, there's a risk for re-emergences to occur, however, if the adaptative process occurred in humans, likely another zoonotic transfer will have to acquire the same adaptive mutations (Andersen et al., 2020).

At the time of writing, the available data does not allow to pinpoint the origin of SARS-CoV-2 (Sallard et al., 2021). Using bioinformatic approaches, it was found that SARS-CoV-2 has a 96.2% identity to the genome of RaTG13 CoV found in *Rhinolophus affinis*. The samples used were sequenced seven years after being collected and the precise location of the collection is not documented, however, a precise match to a fragment from BtCoV/4991 from Yunnan Province was found. This sample was retrieved after the death of three miners from a strange pneumonia (Ge et al., 2016; Sallard et al., 2021; H. Zhou et al., 2020). From faeces of *Rhinolophus malayanus,* also from Yunnan Province, the metagenome of RmYN02, was assembled revealing a 97.2% identity to SARS-CoV-2 (ORF 1ab), corresponding to the first two-thirds of its genome. In contrast,

the other third diverges significantly (P. Zhou et al., 2020). Events of recombination are well described for CoVs and such events can be viewed using Percent of Identical Positions (PIP) profiles (Sallard et al., 2021).

Between RaTG13 and SARS-CoV-2, PIP is higher than 96% through the genome, but decreases to 70% in the S protein (Sallard et al., 2021). This region of SARS-CoV-2 finds its closest sequence in a metagenome from pangolin samples (MP789) (Lam et al., 2020). Thus, the hypothesis that the origin of SARS-CoV-2 is the result of recombination between bat and pangolin infecting viruses arose (Xiao et al., 2020). The most popular hypothesis points to a zoonotic origin, based on phylogenetic studies using the two scenarios described in this topic's first paragraph (Lam et al., 2020; Xiao et al., 2020; P. Zhou et al., 2020). The missing link to strengthen this hypothesis is the lack of the last intermediate host before human infection. This would clarify not only the species, but also the conditions in which the event took place, namely, in wild animals or laboratory ones (Sallard et al., 2021).

Another hypothesis concerning laboratory manufacturing also exists, that postulates either a synthetic origin or a laboratory escape of a chiropteran virus adapted to animal models (Sirotkin & Sirotkin, 2020). The inconclusiveness about SARS-CoV-2 origin makes it impossible to formally rule out any hypothesis here presented; thus, further information is required.

1.5 Surface glycoprotein

The spike glycoprotein (S) is one of the four structural proteins that compose a CoV virion. Recognition by S establishes the cell tropism and host range, since it is a class 1 viral fusion protein that mediates the fusion of both viral and cellular membranes (Bosch et al., 2003; Tortorici & Veesler, 2019). This protein forms homotrimers that protrude through the viral surface, with each protomer consisting of a single polypeptide chain, typically with a minimum of 1100 residues to a maximum of 1600 (Tortorici & Veesler, 2019). S has two functional subunits called S1 and S2. The first constitutes the trimers' apex and includes the receptor-binding domains (RBD), responsible for binding to host cell receptors, it also stabilizes the S2 fusion machinery while in prefusion state. S2 subunit is anchored in the viral membrane and mediates membrane fusion, enabling CoVs entry into the host (Bosch et al., 2003; Tortorici & Veesler, 2019).

Hosts can have an impact on the S protein conformation and, therefore, in viral entry through cleavages. In the case of various CoVs, the host processes the S protein by cleaving it, which originates the aforementioned S1 and S2 subunits, that establish a non-covalent bond. A further cleave, now for all CoVs, is made by host proteases at a location just upstream of the fusion peptide, a location denominated as S2'. By

irreversible conformational modifications, these events are pointed as to activate the protein and allow membrane fusion (Heald-Sargent & Gallagher, 2012; Millet & Whittaker, 2014; Millet & Whittaker, 2015; Tortorici & Veesler, 2019). Therefore, CoVs cell entry is a complex process simultaneously requiring receptor-binding and proteolytic S protein processing (Walls et al., 2020).

Although CoVs use the S1 subunit for receptor recognition, different domains within it might be used depending on the viral species. Species such as the endemics hCoV-OC43 and HKU1 use the domain A (S^A) to attach to 5-N-acetyl-9-O-acetyl-sialosides (Vlasak et al., 1988). MERS-CoV uses the same domain, however the recognised receptor is non-acetylated sialoside (Tortorici et al., 2019). The domain B (S^B) is used by SARS related CoVs, SARS-CoV and SARS-CoV-2, recognizing and directly interacting with angiotensin-converting enzyme 2 (ACE2) (Hoffmann et al., 2020; Kirchdoerfer et al., 2018; Lan et al., 2020; Song et al., 2018). Additionally, SARS-CoV-2, exclusively, possesses a polybasic furin cleavage site located at S1/S2, due to a 12 nucleotide insertion (Walls et al., 2020). This feature distinguishes this species from other SARS-CoVs that present a monobasic cleavage site and may contribute for a wider range of this virus tropism, since the production of furin-like proteases is practically ubiquitous (Steinhauer, 1999). Of the four endemic HCoVs only one conformational state was detected in the S homotrimer (Kirchdoerfer et al., 2016; Li et al., 2019; Tortorici & Veesler, 2019; Walls, Tortorici, Frenz, et al., 2016). In contrast, for SARS-CoV, MERS-CoV and SARS-CoV-2, the opening of S^B at the apex shows different states. The closed state is stable and receptor inaccessible, while the opened state is mentioned as the less stable receptor accessible one. This conformation was proven essential for engaging the cell receptors in these three viruses (Song et al., 2018; Walls et al., 2019). It is also expected that other CoVs that use this domain go through this exposure of the S receptor binding motif (RBM), located at the RBD (Walls et al., 2020; Walls, Tortorici, Frenz, et al., 2016). Once engaged with the cell receptor, the removal of the S1 crown is deemed essential for membrane fusion (Walls et al., 2017). Curiously, endemic HCoVs mostly present S trimers in the closed conformation, while the pathogenic ones possess a partially opened state, a pattern that may be used to identify potential highly pathogenic viruses (Walls et al., 2020).

1.5.1 S protein architecture

The conformational structure of the S1 subunit is a distinguishing factor between the four genera of CoVs. α- and γ-CoVs show a square shape architecture, whilst β- and δ-CoVs present an architecture closely resembling a “V” shape (Shang, Zheng, Yang, Liu, Geng, Tai, et al., 2018; Tortorici et al., 2019; Walls, Tortorici, Bosch, et al., 2016;

Walls, Tortorici, Frenz, et al., 2016). Within S1, the region with the highest variability is the B domain, hence, it also shows different architectures between genera (Tortorici & Veesler, 2019). β-CoVs stand out by mediating receptor engagement with a β-sheet core subdomain adorned by an external variable subdomain, in their B domain (Tortorici & Veesler, 2019; Tortorici et al., 2019). The same domain for α-, γ- and δ-CoVs tackles receptor attachment with a β-sandwich adorned with loops (Tortorici & Veesler, 2019; Walls, Tortorici, Frenz, et al., 2016). Interactions of the B domain while in trimer structure also differ. In α- and δ-CoVs, B domains pack within the same protomer with A domains, while in β- and γ-CoVs, the interactions involve A and B domains of another protomer (Shang, Zheng, Yang, Liu, Geng, Luo, et al., 2018; Walls, Tortorici, Bosch, et al., 2016).

Various α-helices, a β-rich connector domain, an antiparallel core β-sheet, a transmembrane region and a stem helix heading to a heptad-repeat 2 (HR2) form the more conserved S2 subunit (Kirchdoerfer et al., 2016; Shang, Zheng, Yang, Liu, Geng, Luo, et al., 2018; Tortorici et al., 2019; Walls, Tortorici, Bosch, et al., 2016). The conserved nature of S2 extends to a degree of similarity with S proteins of pneumovirus and paramyxovirus, suggesting an evolutionary link between the fusion proteins and the conserved mechanism involved in cell fusion (Tortorici & Veesler, 2019; Walls, Tortorici, Bosch, et al., 2016; Walls et al., 2017). The process by which the virus fuses with the cell is facilitated by specific elements of this subunit, namely, the fusion peptide and the transmembrane domain (Tortorici & Veesler, 2019). In order to complete the membrane fusion process, a trimeric structure needs to be assembled. To ensure a correct trimerization, a conserved tryptophan-rich (Y[VI]KWPW[YW]VWL) portion exists. This segment precedes the transmembrane region, whose interactions within and, perhaps, between the S trimers are considered crucial for complete virus-cell fusion (Schroth-Diez et al., 2000).

1.6 SARS-CoV-2 variants

A Variant of Interest (VOI) can be associated with predicted higher transmissibility or disease severity, having changes concerning receptor binding or negatively impacting the effect of neutralization by antibodies or treatments. Once the predictions become evidence, the variant gets the classification of Variant of Concern (VOC). Additionally, a VOC can produce diagnostic detection failures. By the end of April 2021, no SARS-CoV-2 variant belonged to the third class, a Variant of High Consequence (VHC). An hypothetical VHC would have clear evidence for a significant decrease in preventive measures, a demonstrated inability to diagnose, a recorded high number of reinfections after vaccination or provoking more hospitalizations due to a more severe clinical disease (Centers for Disease Control and Prevention, 2021).

It is estimated that, starting on September 2020, a VOC labelled B.1.1.7 or 20I/501Y.V1, emerged in the United Kingdom (Public Health England, 2020; Rambaut et al., 2020). This variant was characterized by mutations in the S protein, of these, three deletions (Δ69-70 HV and Δ144Y) and a mutation in the RBD, N501Y. The hypothesis of gradual mutations is not supported by the high sequencing practiced in the UK. A possible explanation of the origin of this variant, due to the high number of point mutations, is a prolonged exposure in a patient with immunocompetence. The rational being that in order to withstand the immune response, several immune escape mutations might occur (European Centre for Disease Prevention and Control, 2020b). Of the mutations in the S protein, N501Y is the only reported substitution to confer fitness gains in viral replication and affinity for cell receptor ACE2. These factors combined result in an increase in transmissibility (R(t)) by a factor of 1.35 (Graham et al., 2021; Liu et al., 2021).

One month later, on Nelson Mandela Bay, South Africa, another VOC was detected and called B.1.351 or 20H/501Y.V2 (Tegally et al., 2020). This variant also has the N501Y mutation, but unlike B.1.1.7, it has three mutations in the RBD, two of them in the RBM. The transmissibility rate of this variant was estimated to be 1.5 higher than other previously circulating variants in the area (Pearson et al., 2021). More recently, in January 2021, four travellers from Brazil were tested near Tokyo, Japan, and a new VOC was discovered and labelled as P.1 or 20J/501Y.V3 (Faria et al., 2021). The mutation profile in the RBM of the S protein is identical to B.1.351. Also, in January, a novel variant was described in California and later labelled B.1.429 or 20C/S:452R. This variant had one mutation in the RBM, different from the variants mentioned above (L452R) and the first mutation recorded to become the dominant subtype worldwide, D614G (Isabel et al., 2020). In five months, the sequenced cases with this variant increased from 0 to 50%, exhibiting a transmissibility 20% higher than the wild type (Deng et al., 2021). By the end of January 2021, all variants described above were detected in a large number of countries, with a prevalence of the UK variant (Galloway et al., 2021).

1.7 Angiotensin Converting Enzyme 2

Recognized by the S protein of CoVs such as SARS-CoV and SARS-CoV-2, ACE2 is an integral membrane protein expressed in organs from different systems like the heart, lungs, kidneys and intestines (Donoghue et al., 2000; Yan et al., 2020). Putting aside its link as a viral receptor, ACE2 has a primary biological role in maturating a peptide hormone whose functions are regulating vasoconstriction and blood pressure, called angiotensin. Therefore, this protein has a role in the renin – angiotensin system, indeed, interference in ACE2 expression is strongly associated with cardiovascular diseases (Yan et al., 2020; Yim & Yoo, 2008). Knockout studies in mice revealed no alteration in

blood pressure homeostasis, but revealed the importance for regulating heart function (Crackower et al., 2002). ACE2 also poses as the amino acid transporter B^0AT1's chaperone, allowing it's membrane trafficking (Kowalczuk et al., 2008).

Having 805 amino acids in mammals, ACE2 is comprised of a N-terminal zinc metallopeptidase domain (PD), residues 19 – 611, that cleaves Ang I, yielding Ang-(1-9) (Warner et al., 2004). The C-terminal collectrin-like domain, residues 612 – 740, terminates in a single transmembrane helix (TMH) that prolongs intracellularly in a 40 residue segment. Between PD and the TMH, exists a ferrodoxin-like fold domain (neck) (Yan et al., 2020). A similarity of around 42% exists between ACE2's PD and the corresponding domain of somatic ACE, while the collectrin-like domain of ACE2 presents a 48% identity to collectrin (Towler et al., 2004). Extracellularly, two subdomains comprise the metallopeptidase domain of ACE2 (I and II), displayed in such a way that creates a cleft between them (≈40 Å deep) and are connected by residues 511 – 531 (Towler et al., 2004). This protein can exist in a homodimer conformation mediated in its majority be the neck domain where extensive polar interaction occurs amidst the second and fourth helices.

PD also contributes for interface, however, at a minor level, conferring the ability to transit from closed to opened conformation, in which its subdomain I moves up to 13 Å (Towler et al., 2004; Yan et al., 2020). It is suggested that two trimeric structures of S protein can interact with the ACE2 homodimer. Transmembrane protease serine 2 (TMPRSS2) cleavage at the C terminal segment of ACE2 is linked to an enhanced S protein mediated viral entry (Shulla et al., 2011).

1.7.1 ACE and ACE2

Studies found, using gene structural analysis, that a duplication of a common ancestor gave origin to *ACE* and *ACE2* (Riordan, 2003). The protein ACE2 acts as an ACE antagonist by inactivating angiotensin II, reducing blood pressure, thus ACE2 serves as an ACE inhibitor. Additionally, both genes can be found in *vertebrates*, primitive tunicates and chordates, presenting some evidence of a slow evolutionary rate in early vertebrate speciation (Fournier et al., 2012; Lv et al., 2018). Further, the ratio of nonsynonymous to synonymous substitutions of *ACE* is lower than *ACE2*, suggesting a more conservative evolution of the former, with either having most of their ratios less than 0.1 (Lv et al., 2018). Variations in the gene's structures are also scarce, with *ACE* having mostly 25 exons and *ACE2* 18. However, cases of exon fusion or disruption have been recorded. The conservative nature of both genes is reflected in the consistency of their gene tree with the species tree and only one ortholog for each gene is found through the genome, suggesting an early loss after the whole genome duplication events (Lv et al., 2018).

1.8 Interface: ACE2 – SARS-CoV / 2 S protein

As stated above, SARS-CoV and SARS-CoV-2 bind to ACE2 using the RBD of their S protein. HeLa cells not expressing ACE2 are not susceptible to those viruses, while those who express are. Moreover, it is known that the RBD binds to ACE2 with a low nanomolar affinity, further proving the importance of ACE2 for infection (Lan et al., 2020). The RBD of SARS-CoV-2 is composed of a twisted five-stranded antiparallel β sheet (β1-4 and 7). The receptor binding motif (RBM) is an extended portion of RBD that contains β5 and β6 strands, loops and α4 and α5 helices located between β4 and β7, said to have the most contacting residues that bind to ACE2 (Lan et al., 2020).

Comparing the overall binding method used by SARS-CoV and SARS-CoV-2, little differences can by pointed. The buried surface is near identical, with SARS-CoV-2 achieving a total of 1687 $Å^2$, only 12 less than SARS-CoV (Lan et al., 2020). The similarities expand to the contacting residues, with 20 ACE2 residues contacting with 17 SARS-CoV-2 and 16 SARS-CoV RBD residues. Out of the 20 ACE2, 17 interact with both RBDs. Analysing the RBDs revealed that 14 residues belonging to RBM were shared between both viruses, with 8 having identical residues, 5 sharing similar biochemical properties and the last one interacting with the same 5 ACE2 residues (Lan et al., 2020).

Outside the RBM, SARS-CoV-2 possesses a unique interaction in residue 417, not present in SARS-CoV, possibly due to the change from lysine to valine. SARS-CoV-2 RBD also has a positive charge electrostatic potential due to Lys 417 that is not present in SARS-CoV. There are also 13 hydrogen bonds and 2 salt bridges in SARS-CoV-2 RBD – ACE2 interface, one less salt bridge than SARS-CoV (Lan et al., 2020). All these differences contribute to a binding affinity of SARS-CoV-2 RBD – ACE2 of 4.7 nM that differs from SARS-CoV's 31 nM. The list of interacting residues goes as follows for SARS-CoV-2 RBD: K417, G446, Y449, Y453, L455, F456, A475, F486, N487, Y489, Q493, G496, Q498, T500, N501, G502 and Y505. For ACE2: Q24, T27, F28, D30, K31, H34, E35, E37, D38, Y41, Q42, L79, M82, Y83, N330, K353, G354, D355, R357 and R393 (Lan et al., 2020).

1.9 Protein – Protein docking

1.9.1 Fully automated computer algorithms

When determining the biological function of a given protein, it is essential to understand its three-dimensional structure, given that it's the 3D shape that dictates how proteins interact with ligands and other proteins (López et al., 2007; Zheng et al., 2019). The process by which protein's 3D structures are experimentally determined varies, in the early 2000s there were two main methods: the nuclear magnetic resonance (NMR) and the X-Ray crystallography. However, consultation of the protein data bank statistics (https://www.rcsb.org/), reveals that after 2017, Electron microscopy is now the second most used method, second only to X-ray. Due to the rise of omics, sequencing has become much faster and cheaper, resulting in large amounts of sequences being produced every day. As of February 2020, more than 207 million protein sequences are deposited in UniProt ("UniProt: the universal protein knowledge base," 2017). A huge number, especially when compared to the modest 150 thousand protein 3D models deposited in Protein Data Bank (PDB) (Berman et al., 2003). To respond to the high amount of sequence data and allow the use from non-experts, fully automated computer algorithms such as Iterative Threading Assembly Refinement (I-TASSER) algorithm have emerged (Yang et al., 2015; Zhang, 2008; Zheng et al., 2019). As an example of these algorithms, I-TASSER, considered in the Critical Assessment of Structure Prediction experiments as the top ranked method for 3D protein structure prediction, consists of a pipeline in a web server that accepts FASTA files or amino acid sequences and produces the five best 3D models (Moult et al., 2016; Moult et al., 2018; Zhang, 2008). Algorithms such as this, use information available of template proteins with potentially similar structures to the protein in question. However, for regions where there is no template available, like in alignment gaps, the structure is built *ab initio* by the algorithm (Zheng et al., 2019). For I-TASSER, typically, results arrive within 1-2 days (Zheng et al., 2019).

Other information can be expected from the output, such as predicted structural features and similar proteins to the query (Zhang, 2008; Zheng et al., 2019).

1.9.2 Docking algorithms

As stated above, there are more than 150 thousand protein structures deposited in PDB, while the number of protein complexes hasn't reached the 18 000 mark (Berman et al., 2003). A few difficulties arise when studying protein complexes, namely the lack of high-level information at an atomic level (Dominguez et al., 2003). Additionally, X-ray technique, although the most used, also faces challenges, having crystallization effected by the complex formation dynamics. NMR, on the other hand, excels in mapping protein-

protein interfaces and the amino acids involved in complex formation, but lacks in providing data about the relative orientation of the involved proteins (Dominguez et al., 2003; Zuiderweg, 2002).

Given the restrains of the described experimental techniques, theoretical methods, i.e., algorithms, based on protein docking started to emerge. The basic principle of said programs involved *ab initio* docking of protein, where one of the molecules remained static, while the other moved through rotations and/or translations (Dominguez et al., 2003; Smith & Sternberg, 2002). Each new configuration was then scored based on parameters such as van der Waals repulsion. As one can tell, these approaches have the drawback of being calculation heavy, since it requires searching through the entire conformational space of the complex. Additionally, these techniques operating in the absence of experimental restrains revealed a poor accuracy (Dominguez et al., 2003).

To solve this problem, the algorithms started to integrate biochemical, bioinformatic and biophysical interaction data, such as the case of High Ambiguity Driven protein-protein Docking (HADDOCK), a data-driven docking approach for modelling biomolecular complexes. Integrating information from interface restrains from NMR, X-ray scattering, electron microscopy or mutagenic experiments, among others, contributes to the high fidelity of this method (Dominguez et al., 2003). Along with protein-protein docking, it is also possible to model single molecule types such as DNA and RNA and protein-RNA/DNA mixed complexes (van Zundert et al., 2016).

1.10 Positively selected sites (PSS)

A region with no variation may be a critical region for the protein in question. Regions with variation, however, may be adaptive (or neutral). Indeed, variations in the amino acid sequence of a protein may be linked to a change or gain of function. These can vary from a slight increase in ligand affinity to the acquisition of resistance to an antibiotic. As is known, not all polymorphisms result in amino acid changes, such as the case of synonymous substitutions. One method to determine the presence or absence of positive selection is to evaluate the number of nonsynonymous (d_N) and synonymous (d_S) substitutions per site (Nielsen, 2001). If their coefficient (d_N/d_S) is equal to 1, then selection is of no importance. For positive selection the coefficient has to be >1, while purifying selection is characterized by a coefficient <1. Purifying selection is expected to be found in functional genes, where conserved regions are frequent (Nielsen, 2001).

Eleven months after being considered a pandemic, SARS-CoV-2 has been found to be under weak purifying selection (MacLean et al., 2021).

2. Objectives

Due to SARS-CoV-2 pandemic impact and the cross-species transmission phenomenon associated with its origin, it is fundamental to know how to identify other potential vectors for SARS-CoV-2. It is also important to assess new emerging variants of SARS-CoV-2 and prevent future transmissions from other species of coronavirus. Our focus will be on the interface between the virus S protein and ACE2, involved in viral cell entry. This study intends to create a simple *in silico* methodology, tackling these issues. For that, I will:

1- Assess the transmissibility rate of new emerging SARS-CoV-2 variants, by estimating the number of interfacing residues with hACE2. Measure the impact of single point mutations, observed in the S protein, in protein binding strength with hACE2.

2- Develop a predicting model able to reliably pinpoint potential SARS-CoV-2 hosts, using an *in silico* approach, by i) establishing a reference value for protein binding strength – yielding candidates where it is expected to find infection; ii) understand infection patterns in regards to specific protein residues or their phylogenetic relations; iii) test ACE2 as essential for SARS-CoV-2 infection in animals; iv) validate the use of simpler forms of surface glycoprotein in the prediction model.

3- Infer the transmission threat level of species belonging to *Coronaviridae*, by: i) determining the presence of essential amino acids involved in the recognition of *Homo sapiens*' ACE2 (hACE2); ii) determine a sequence pattern able to distinguish species by their ability to bind to ACE2; iii) obtain knowledge on amino acid positions under positive selection in different *Coronaviridae* species, indicative of possible adaptations to the host.

The developed models will provide new insight on SARS-CoV-2 infection and are also a quick and inexpensive way to evaluate present and future threats in regards to *Coronaviridae* infections.

3. Methods and materials

3.1 Sequence retrieval

3.1.1 ACE2 and reference S protein

The acquisition of ACE2 sequences started by querying for “Animals” in the “Assembly” option of NCBI. All genomes in the RefSeq database were downloaded as CDS from genomic FASTA (*.fna*). As for SARS-CoV-2 S protein (YP_009724390.1), its sequence was extracted from the reference genome deposited in NCBI (NC_045512.2). The animal genomes extracted were processed in order to retrieve the ACE2 sequences using the *Preparing datasets for large scale phylogenetic analyses* protocol (https://www.sing-group.org/seda/manual/index.html) and the dataset builder software SEDA. To do so, a *tblastn* was performed using each genome separately, the hACE2 (NP_001358344.1) as query, and 0.05 as expected value.

Then, taxonomic information was added to the headers and the files merged into a single document. After reformatting, the sequence headers contained the species name, species common name, class, family and accession number. Since the sequences encode for a protein, only complete data is of interest, therefore, all sequences containing any ambiguous nucleotides, in frame stop codons, non-multiple of three sequence length or a start codon different from ATG were excluded. To further refine the dataset acquired and ensure the absence of annotation errors, sequences with more than 10% difference in length to hACE2 were also removed. Lastly, to avoid unnecessary information, redundant sequences and isoforms were discarded. To control any loss of information, the *Homo sapiens*’ sequences were monitored and used as reference.

I) <u>Infected and non-infected species</u>

Creating a perdition model requires the use of controls, therefore, two separate list were made. The first list was composed of species where infection by SARS-CoV-2 was observed *in vivo* or *in vitro* representing the positive control (Gryseels et al., 2020; Gu et al., 2020; Heegaard et al., 2020). The species present were *Homo sapiens, Felis catus, Macaca fascicularis, Macaca mulatta, Mustela putorius furo, Peromyscus maniculatus, Routettus aegyptiacus, Mesocricetus auratus, Callithix jacchus, Panthera leo, Sus scrofa*, *Neovison vison* and *Mus musculus.* The second one consisted of species where SARS-CoV-2 infection was absent in *in vivo* or *in vitro* studies, the negative control

(Suarez et al., 2020). Those being: *Anas platyrhynchos domesticus, Gallus gallus, Anser cygnoides, Coturnix japonica* and *Meleagris gallopavo f. domestica*. It is important to refer that *Sus scrofa* has been proven to not be infected *in vivo*, however, experiments show that HeLA cells expressing pig's ACE2 were susceptible to infection, hence, for the purposes of this study, it will be considered as a susceptible species (Gryseels et al., 2021).

The next step involved the acquisition of the ACE2 amino acid sequences of these two lists. If a species was present in the ACE2 dataset, the sequence was retrieved from NCBI using the accession number present in the header, if not, the sequences used were retrieved from the *Ensembl* database by searching for *hACE2* (ENSG00000130234) orthologues.

3.1.2 *Coronaviridae* S proteins

To gather sequences from *Coronaviridae* S protein from NCBI, the "Assembly" option was queried using "*Coronaviridae*". Sequences from RefSeq as well as GenBank were retrieved. An approach identical to the one described for ACE2 sequence retrieval was applied, using SARS-CoV-2 S protein (YP_009724390.1) as query for *tblastn* step.

To complement NCBI, ViPR, a more complete database for virus information, was used. Searching in *Coronaviridae* family and querying for "spike glycoprotein" all sequences associated with complete genome annotation were downloaded. Using SEDA, the dataset was filtered to exclude sequences with redundant nucleotides and sequences, in frame stop codons, non-multiple of three sequence length, non ATG start codon, isoforms and sequences with more than 10% sequence length difference from the 1273 amino acid SARS-CoV-2's S protein. No taxonomic information was added to the sequence headers as ViPR's files don't contain this metadata.

3.1.3 SARS-CoV-2 variants' S protein

For the acquisition of these sequences, their set of mutations were retrieved, preferentially, from pangolin (https://cov-lineages.org/resources/pangolin.html), complemented, when necessary, with CDC (Centers for Disease Control and Prevention, 2021; O'Toole et al., 2021) (Table 1). The set of mutations were individually edited in the FASTA file containing SARS-CoV-2 reference S protein (YP_009724390.1). Mutations belonging to more than one variant or unique in a VOC were selected to be added as point mutations to the reference S protein.

Table 1 – List of VOCs and VOIs identified by their Pango lineage identifier and their respective S proteins mutations.

Variants	Mutations
A.23.1	F157L; V367F; Q613H; P681R *
B.1.1.7	del69-70HV; del144Y; N501Y; A570D; P681H; T716I; S982A; D1118H *
B.1.351	D80A; D215G; K417N; E484K; N501Y; A701V *
B.1.427	L452R; D614G **
B.1.429	S13I; W152C; L452R; D614G **
B.1.525	A67V; del69-70 HV; del144 Y; E484K; D614G; Q677H; F888L **
B.1.526	L5F; T95I; D253G; S477N; E484K; D614G; A701V **
P.1	L18F; T20N; P26S; D138Y; R190S; K417T; E484K; N501Y; H655Y; T1027I *
P.2	E484K; D614G; V1176F **

* According to O'Toole et al. (2021)

** According to Centers for Disease Control and Prevention (2021)

3.2 Phylogenetic approach

3.2.1 ACE2 gene tree

Using as input the sequences obtained previously from NCBI for ACE2, a phylogenetic approach was used. This enabled the visualization of ACE2 gene tree, revealing cases of gene duplication or deletion in the dataset. As a way to speed the process and given the large number of sequences, the dataset was aligned using MUSCLE in MEGAX. Phylogenetic trees using the nucleotide and amino acid sequences were made by using the Neighbor-joining statistical method with a bootstrap value of 500 and maximum composite likelihood as model.

3.2.2 S protein gene tree

An identical approach was made regarding the S protein sequences. The dataset retrieved from NCBI and ViPR were also used to produce phylogenetic trees in MEGAX. This served to assess the 10 species with the S protein more closely related to SARS-

CoV-2. For the same purpose a *blastp* of the ViPR dataset was also conducted using as query SARS-CoV-2's S protein. Closely related species were selected for protein docking.

3.2.3 Consensus sequences

I) ACE2 consensus

Since one of the objects of study is the ACE2 protein, it became important to identify variable and conserved sites across the acquired animal species. Substitutions on the nucleotide level may not result in amino acid change, hence, for clearly defined clusters present in the ACE2 gene tree (bootstrap >80), the sequences were translated, aligned using Clustal Omega and a consensus sequenced constructed. The output file had the verbose notation showing all the amino acids for a given position with a frequency above 20%.

II) SARS-CoV-2 S protein consensus

Viruses are very prone to mutations in their genome. This is further attested by the emerging new variants of SARS-CoV-2 only one year after the first recording in humans. These mutations may carry important modifications to the way the S protein interacts with ACE2, and so the need to construct a consensus sequence of the protein arose. To do so, the gene name "Surface glycoprotein" was used as query in ViPR as of the 26th October 2020. At the time, the database contained 33723 strains. All results, including sequences from incomplete genomes were downloaded and processed similarly to *Coronaviridae*'s S proteins to remove contaminant sequences, i.e., sequences with ambiguous annotations or lengths non-multiple of three. The filtered data was then translated, aligned using Clustal Omega and made into a consensus sequence in verbose mode, showing all existing variations.

The whole process following the sequence retrieval and culminating in consensus sequences was performed using SEDA.

3.3 Protein-protein docking

3.3.1 Protein 3D structure prediction

Predicting protein-protein interactions requires the use of 3D models instead of amino acid sequences. Therefore, 3D protein structure prediction was made using the *Iterative Threading ASSEmbly Refinement* (I-TASSER) web server (http://zhanglab/I-TASSER/) for the ACE2 of the positive and negative control groups, the S protein of SARS-CoV-2, its VOCs, VOIs and point mutations, as well as, of phylogenetically close

species to SARS-CoV-2. The output files consist of 5 prediction models per amino acid sequence given as input. To pick the model to be used in further steps, a 3D structure for hACE2 (1R42 - DOI: 10.2210/pdb1R42/pdb) and SARS-CoV-2 (6VYB – DOI: 10.2210/pdb6vyb/pdb) S protein were obtained from the Protein Data Base (PDB; http://www.rcsb.org/pdb/).

For SARS-CoV-2 S protein, the 3D model in PDB represented the protein in an asymmetric homotrimeric structure. Two chains in the closed and one in the opened conformation. As mentioned above, SARS-CoV-2 S protein binds to hACE2 by the chain in the opened state, therefore chain B was isolated using PyMOL.

A TM-align (Zhang & Skolnick, 2005) for all the I-TASSER outputs was then made. The 5 predicted models for each ACE2 protein were aligned with 1R42. The same method was applied for SARS-CoV-2 variants' S protein using chain B of 6VYB. The model yielding the highest TM-Score ($0,5 \leq x \leq 1$) for any given case was chosen for further analysis.

3.3.2 HADDOCK

The protein-protein docking prediction was made using the HADDOCK2.2 (https://alcazar./HADDOCK2.2/) web server. The algorithm requires as input two protein structures as well as a list of active and passive (optional) sites for each protein. Residues showing a high solvent accessibility in the free form molecule and a significant chemical perturbation after complex formation in NMR titration data are classified as active. As for passive residues, they are defined by having a lower chemical perturbation and/or directly flank surface active residues while also having high solvent accessibility (Dominguez et al., 2003).

The structures submitted consisted of the I-TASSER model selected by using the TM-align method. For good practice the 1st molecule used was the non-variable one. Additionally, Lan et al. (2020) provided a list of SARS-CoV-2 and hACE2 contacting residues.

All amino acid sequences of ACE2 used were aligned in MEGAX using the MUSCLE algorithm and the ortholog sites corresponding to hACE2 contacting residues were identified and used as active residues. The same procedure was conducted for all S proteins. For added flexibility and confidence in the results yielded by HADDOCK, the 3 left and right flanking positions of a contacting site were also added to the active residues list. The passive residues were defined automatically around the active ones. The four best structures from the cluster with the lowest Z-score, given as output, were considered for further analysis. Z-score represents the number of standard deviations the HADDOCK score of a cluster is to the mean of all clusters, therefore, the lower the Z-score,

the better (van Zundert et al., 2016). Each S protein-ACE2 docking, yields four predicted structures.

Three groups of data were defined: one consisted of the docking of SARS-CoV-2 variant's S protein with the hACE2 (V-hACE2), another of SARS-CoV-2's S protein with ACE2 of the positive and negative controls for infection (SARS-CoV-2-ACE2) and the last consisted of *Coronaviridae* species docked with hACE2 (*CoVs*-hACE2).

3.3.3 PDBePISA

To enable the visualization of interacting residues between the various S protein and ACE2s, the data retrieved from the HADDOCK step was submitted to PDBePISA (https://www.ebi.ac.uk/pdbe/pisa/). For each case, the four best structures were submitted and the interface data viewed.

The docking structure with the most interfacing residues in the S protein (1st molecule) or in the ACE2 (2nd molecule, as a tie breaker) was picked and it's *Interfacing residues* tables transferred to a worksheet. An interfacing residue is determined by its buried surface area (BSA column) >0. Hence, this column was converted in a binary result, 1 for interfacing, 0 for non-interfacing. For the three groups of data, the S protein and ACE2 tables were grouped separately, with the results presented according to the amino acid sequence alignment, when needed.

The percentage at which residues interact in SARS-CoV-2 – hACE2 data group was calculated and a list composed of core residues for interaction for both the S protein and ACE2 ($f \geq 90\%$) was made. Additionally, for the data group SARS-CoV-2 – ACE2, a similar list was made only for the positive control. Furthermore, both lists containing the core residues for interaction were intersected in order to obtain the sites with the highest interfacing frequency.

3.4 Large Scale predictive model

The protocol here described, together with the predictive model involving the core residues, although fast, it has some steps that need to be performed one species at a time. So, the next logical step involved the creation of a large scale implementation of the SARS-CoV-2 susceptible species predictive model. The amino acid composition of the positive control and SARS-CoV-2 variants' S protein groups for SARS-CoV-2 infection was assessed for the intersected ACE2 core residues. Only amino acids that never appeared as non-interfacing were considered for that given position. Conserved residues were treated as such, having only one valid option, non-conserved sites were grouped,

when possible, by a common characteristic, such as size, charge or solubility. In cases where grouping was unavailable, the site was also discarded.

The data regarding ACE2 sequences obtained in the *Sequence retrieval* step were translated and aligned in MEGAX using MUSCLE algorithm. Then, the core residues positions were identified using hACE2 as reference. For all species present, a score of 1 was attributed if a core site contained a valid amino acid, determined by the method described above, or a 0 otherwise. Only species with a perfect score were considered prime candidates for SARS-CoV-2 infection.

3.5 S protein: Sites under positive selection

The sequences used for the determination of positively selected sites (PSS) were extracted from the *Coronaviridae* database of ViPR. By querying for the gene name "S", all sequences, including ones belonging to incomplete genome annotations were gathered as FASTA files. The sequences were then grouped by species, and the sets used as input for the Integrated Positively Selected Sites Analyses (IPSSA[1]) Compi Pipeline. IPSSA allows the analysis of PSS from large data sets using two phylogenetic, FUBAR (Murrell et al., 2013) and codeML (Yang, 1997) and one population approaches, OMEGAMAP (Wilson & McVean, 2006). This characteristic circumvents the disadvantage of each individual approach. The pipeline uses FASTA files of CDSs as input, removing stop codons and line breaks. Sequences with ambiguous positions and lengths non-multiple of three are searched for and removed. Sequences are used to produce an amino acid level master alignment using the T-coffee suite (Denoeud et al., 2014).

Apart from the FASTA sequences, the codeML and FUBAR approaches require a phylogenetic tree to be given as input. For this, IPSSA uses MrBayes (Ronquist et al., 2012). Additionally, to prevent false positives of PSS due to recombination, the pipeline also runs Phipack[2] software. As output, the IPSSA returns the PSS for every method applied, the alignment gaps of the master alignment, evidence for recombination and sites with low alignment support value (López-Fernández et al., 2021).

For speed and processing power purposes due to the acquisition of Bayesian trees, subsets of sequences were created. If a dataset given as input had less than 60 sequences, the pipeline would run 1x using all available data. If the number exceeded 60, but was lower than 120, four random subsets of 60 sequences were created and given as input for the pipeline. A similar approach was made for species with a total number of sequences not exceeding 240, but with the formation of eight subsets instead of four.

[1] https://sing-group.org/compihub/explore/5fa91806407682001ad3a1e9

[2] https://www.maths.otago.ac.nz/~dbryant/software/PhiPack.tar

For SARS-CoV-2, the sequences were divided by year of publication given their number surpassing one hundred thousand. Then, 100 subsets were formed containing all sequences from the year 2019, 60 sequences from 2020 and 60 sequences from 2021. From these subsets, 60 random sequences were chosen to be used as input for the IPSSA pipeline. For any sequence set or subset with a K_s higher than 0.1, the alignment was checked to ensure only sequences belonging to the same species were present. Any significantly contrasting sequence was removed from the sets and subsets.

From the output, PSSs from FUBAR and codeML were considered valid. The ones given by OMEGAMAP were discarded if the output consisted of several consecutive sites. The positions given by IPSSA as PSSs correspond to the positions of the master alignment of the set or subset in question, not the actual sites in the amino acid sequence. Thus, a single sequence that belonged to each master alignment was aligned with the reference S protein of SARS-CoV-2. The PSSs were searched in these alignments to determine the ortholog site in SARS-CoV-2.

3.6 *Coronaviridae*'s use of ACE2

The sequences of species with closely related S proteins to SARS-CoV-2, determined by *blastp*, were aligned to enable the search of sequence patterns related to the use of ACE2. Additionally, the protocol for protein – protein docking was also applied to those species, with the aim to assess the interfacing nature of established core residues for both the S protein and hACE2. Simultaneously, the models chosen, obtained from I-TASSER, where aligned with SARS-CoV-2 and trimmed to only contain their RBM (Lan et al., 2020). Furthermore, bibliography was queried to acquire species where the use of ACE2 for cell entry, whether positive or negative, was confirmed and tested. The RBMs of the closely related S proteins and from the ones retrieved from the bibliography were then aligned with SARS-CoV-2 reference RBM using TM-align and the score normalized for the reference size was annotated.

4. Results

4.1 ACE2 phylogeny

The ACE2 sequence retrieval from NCBI yielded, after SEDA's *tblastn* 3374 sequences. For *Homo sapiens*, 15 sequences were identified as ACE1 and ACE2 isoforms, as well as, collectrin. The latter is a glycoprotein that shares homology to the C-terminal portion of ACE2 (Zhang et al., 2001). During SEDA's processing, of the 15 human sequences, 4 were discarded for their redundancy, 7 had more than 10% size difference

to the 805 amino acid reference length and 2 were considered as isoforms, leaving only 2 sequences (≈13%) on the dataset: ACE1 isoform 1 (XP_006721800.2) and ACE2 isoform 1 (NP_001358344.1). By the end of the sequence filtering step, a total of 600 sequences remained (≈18%).

A gene tree was obtained in MEGAX using all sequences and the Neighbor-Joining method with the Anthozoa cluster used as root. Class wise, the nucleotide sequence was able to form clear clusters, with the exception of Insecta, which appears scattered in various clusters. Actinopterygii and Mammalia appear twice, suggesting a duplication of the ACE2 gene. Further analysis revealed that the Mammalia and Actinopterygii cluster has ACE1 gene sequences (Fig.1 – blue cluster). The biggest cluster found with a bootstrap of 99 ranges from Mammalia to Hyperoartia class (Fig.1 – orange cluster). Since it includes the ACE2 gene sequences, it will be referred as the ACE2 cluster and is represented in detail in Fig. 2.

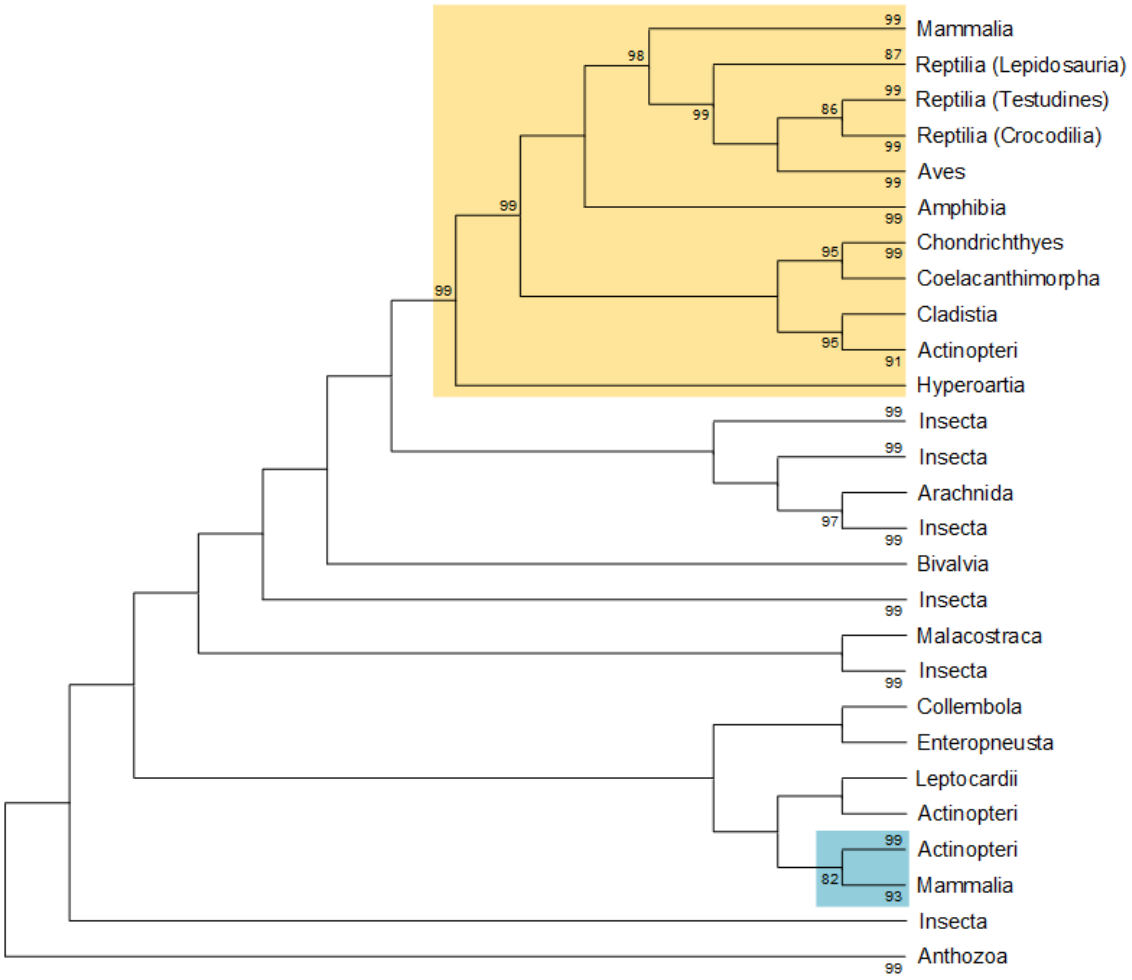


Fig.1 – Neighbor-Joining ACE2 protein tree cladogram, with Anthozoa cluster used as root, collapsed by classes. Bootstraps values of 500 replicates are presented next to the branches, values under 80 are not shown. The yellow cluster represents the mammalian cluster with ACE2 sequences. The blue cluster depicts the mammalian cluster with ACE1 sequences. A total of 600 protein sequences were analysed in the final dataset conducted by MEGAX.

The ACE2 cluster is composed of 290 sequences, with 120 belonging to mammals, following the accepted species tree (Meyer & Zardoya, 2003).

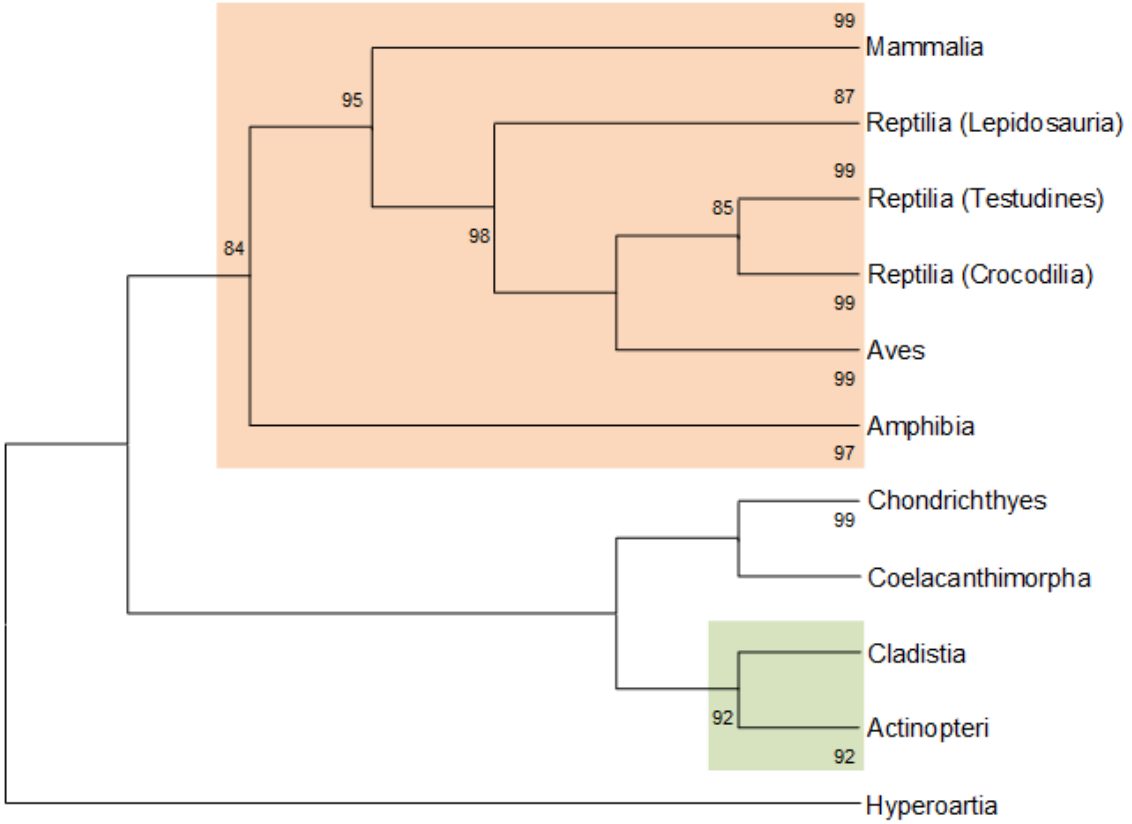


Fig.2 – Cladogram of the bootstrap consensus sequence ACE2 amino acid tree by Neighbor-Joining method, collapsed by classes, with Hyperoartia as root. The bootstrap value of 500 replicates is presented, bootstrap values under 80 are not shown. Mammalian cluster is represented in orange. A total of 290 aa sequences were analysed conducted by MEGAX.

4.2 Consensus sequences

4.2.1 ACE2

The consensus sequence produced with the 600 ACE2 sequences retrieved from NCBI was analysed, particularly the contacting sites of hACE2 with SARS-CoV-2 S protein (Lan et al., 2020). Of the 20 sites, 12 are conserved in at least 80% of the sample. From the remaining residues, only sites 34, 79 and 354 presented more than two amino acids with a frequency higher than 20%. It is of note that site 34 was the most variable, given that no amino acid holds a frequency of over 20%.

4.2.2 SARS-CoV-2 S protein

Processing the ViPR data regarding SARS-CoV-2 strains yielded 975 sequences. The consensus sequence of this dataset shows that variations in the protein

sequence exist, but at a small frequency. This becomes clear as only two positions (477 and 614) present more than one amino acid with a frequency above 5%. However, if every variation with a frequency above 0% is considered, the number of variable positions would surpass the conserved ones and alignment gaps would also be present.

4.3 SARS-CoV-2 variants

4.3.1 Number of interfacing residues

To evaluate the impact of protein binding affinity on the transmissibility rate of SARS-CoV-2 variants, the number of interfacing residues shared between hACE2 and the S protein was determined. The protein-protein interaction of SARS-CoV-2 variants' S protein with hACE2 shows that all variants presented a higher number of total interfacing residues than the reference S protein, the latter with a score of 71 residues (Fig.3). In contrast, the highest number recorded of 98, belongs to the variant of concern B.1.351, which represents an increase of 38%. On average, the number of interfacing residues is 75.3 and 81.4 for variants of interest and concern, respectively, however, no statistical differences were detected using the non-parametric Mann-Whitney U test (*P*>0.05).

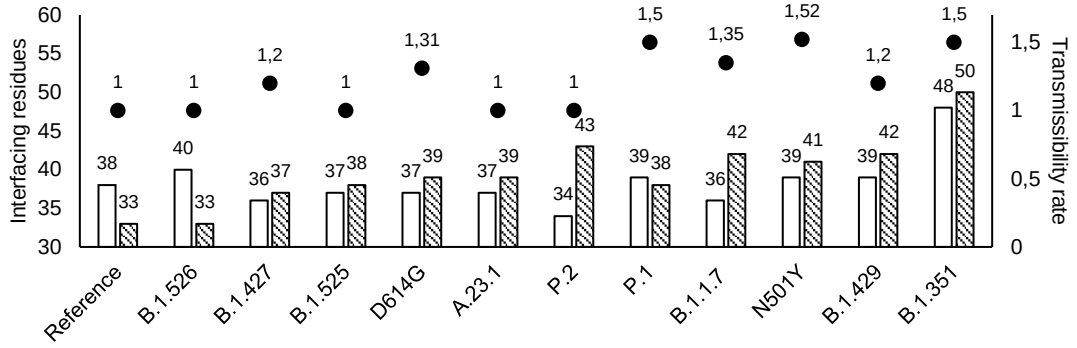


Fig.3 – Number of interfacing residues of SARS-CoV-2 variants' S protein (white) and hACE2 (stripes), with their respective documented or expected transmissibility rate (dots). Results ordered by total number of interfacing residues. Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues.

Using the expected transmissibility rate documented and the total number of residues in interaction, using the non-parametric Spearman's ρ test, a correlation coefficient of 0.35 (*P*=0.044) was found. Focussing on the interfacing residues of hACE2 and S protein individually, no correlation was found between S protein's interfacing residues and the transmissibility rate of SARS-CoV-2 variants (Spearman's ρ test; *P*>0.05). In opposition, the ACE2's interfacing residues of the sample positively correlate with the transmissibility rate with a R=0.768 (Spearman's ρ test; *P*=0.004) (Fig.4).

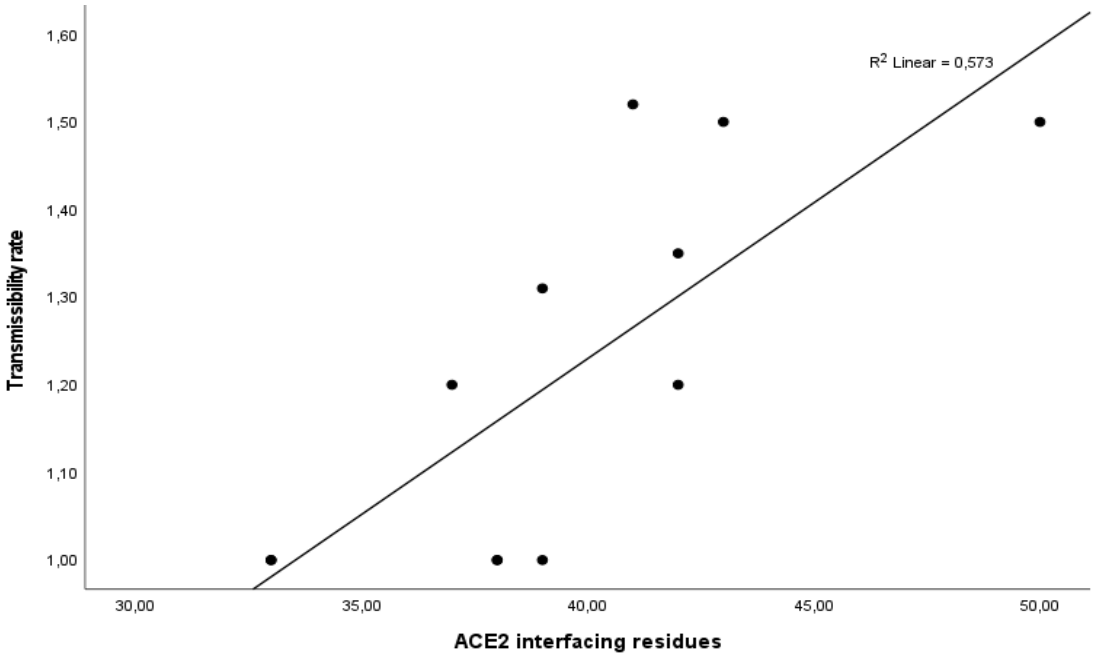


Fig.4 – Correlation between the number of interfacing residues of SARS-CoV-2 variants and their expected and documented transmissibility rate. Image obtained using SPSS.

4.3.2 Hydrogen bonds and Salt bridges

To further study the impact of protein binding strength in variant transmissibility, the quantification and location of bonds such as salt bridges and hydrogen bonds, were determined. Despite the high binding affinity to hACE2 attributed to SARS-CoV-2, no S protein variants yielded disulfide bonds, presenting only hydrogen bonds and salt bridges. Amidst the two, the former presents, across all S proteins, a total of 144 bonds, far exceeding the 50 of the latter (Supplementary Table 5).

The variant of interest P.2 is of particular interest, since it presents the second lowest number of hydrogen bonds, the same as reference, however it is the only case lacking salt bridges. B.1.429 presents the highest number of bonds in the sample, 21 (Fig.5). Supplementary Table 5 for the S protein shows that residues 455, 456 and 490, represented as interfacing for 100% of the variants, never establish a hydrogen bond or salt bridge. There is no feature able to distinguish the reference bond pattern from the other variants or VOIs from VOCs. The average number of salt bridges, hydrogen bonds and their sum show no statistical significance between variants of concern and of interest (non-parametric Mann-Whitney U test; $P>0.05$).

There is, however, a higher prevalence of hydrogen bonds in residues 484-505 in the S protein of the variants, particularly in residues 484, 493, 494 and 505. Salt bridges are only found after residue 417 in variants and only variants of concern have them in residue 484. Most of the S proteins form hydrogen bonds with residues 30, 38 and 353 of hACE2. It is of note that all S proteins possess hydrogen bonds in at least one of these three sites. Compared to reference, in general, variants present more

hydrogen bonds in residues between the positions 42-393. Salt bridges are concentrated in residues 30 and 38 and, although scarce, after residue 38, are only found in SARS-CoV-2 variants.

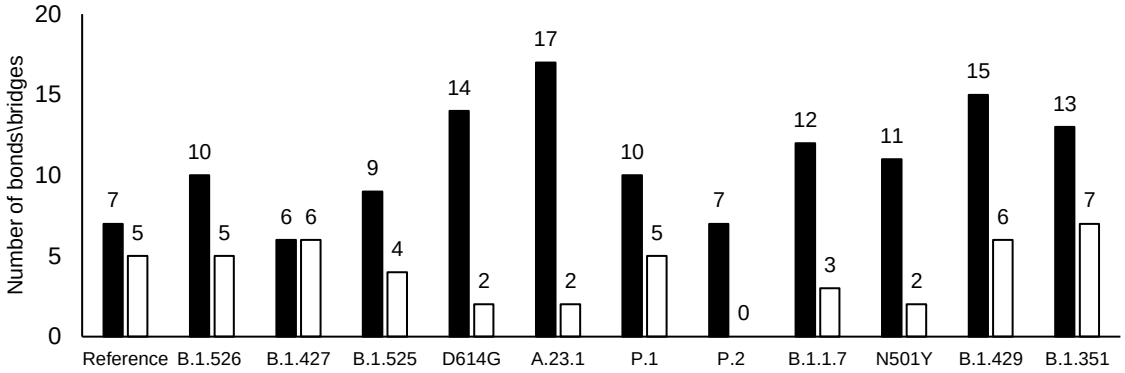


Fig.5 – Number of hydrogen bonds (black) and salt bridges (white) formed by the docking of hACE2 and the S protein of variants of concern and interest and with single mutations that have a documented or expected transmissibility rate. Results ordered by total number of interfacing residues. Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues.

4.3.3 Point mutations

A variant is defined by a group of mutations. The impact of a single mutation may be enhanced or inhibited by another, hence it became important to isolate recurring mutations across variants and evaluate their individual impact. Given that only ACE2 interfacing residues presented a correlation with transmissibility rates, they will be the focus.

Variants with the N501Y mutation always present a total of ACE2 interfacing residues higher than that of the wild type (Fig.6A). When isolated, this mutation exhibits more interfacing residues (41) less than P.1, B.1.1.7 and by B.1.35, all with N501Y as a shared mutation, Furthermore, mutating the asparagine to an aspartic acid instead of a tyrosine yields greater results.

The mutation D614G, when present, always present a total of ACE2 interfacing residues equal or higher than that of the wild type (Fig.6B). While isolated, increases the number of interfacing residues by 6, from 33 to 39, More than B.1.525, B.1.427, B.1.526 and P.2. However, if this mutation is added to the variant B.1.1.7 or P.1, the ACE2 interfacing residues decreases, but aren't lower than the isolated mutation. Adding the mutation D614G to N501Y yields a result that is, simultaneously, higher than D614G and lower than N501Y.

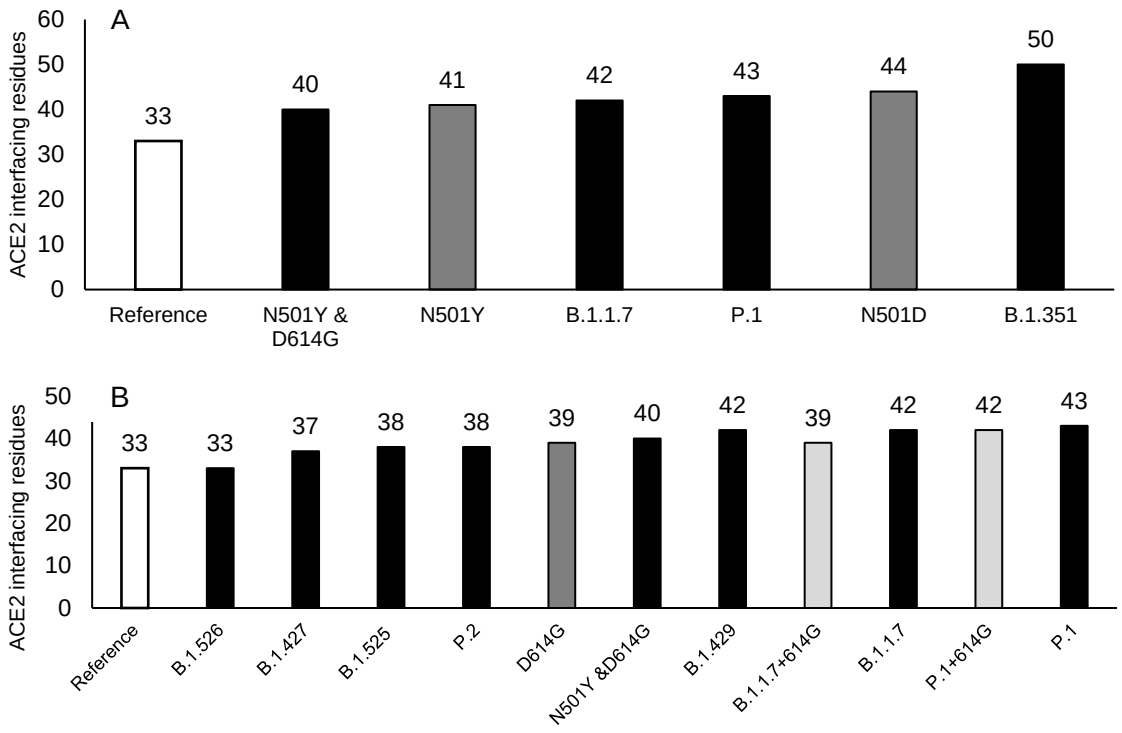


Fig.6 – Effect of the point mutations N501Y (A) and D614G (B) to the ACE2 number of interfacing residues between hACE2 and S protein when applied to the reference sequence (grey), in tandem with variant defining groups of mutations (black) or added to existing variants (light grey). Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues.

All variants with the point mutation L452R present a number of ACE2 interfacing residues higher than wild type (Fig.7A). While isolated, this mutation shows more interfacing residues than VOC B.1.427. In contrast, the point mutation E484K isn't always associated with a higher number of interfacing ACE2 residues, such as the case of VOI B.1.526 (Fig.7B). This variant shows the same result as the wild type. The isolated mutation, however, allows a significant increase in interfacing residues (46), higher than N501Y, D614G and L452R. Only surpassed when in tandem with the other mutations of the variant B.1.351. The VOI B.1.526, B.1.525 and P.2 present a significant lower result when compared to the isolated mutation, despite having this mutation.

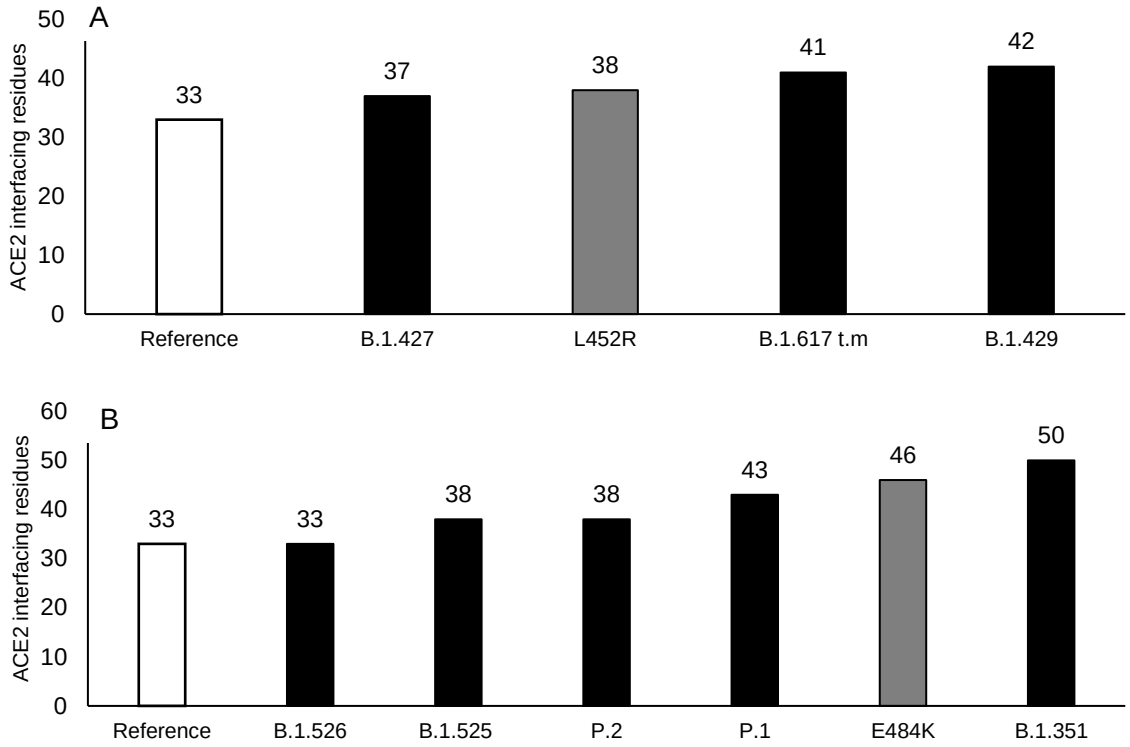


Fig.7 – Effect of the point mutations L452R (A) and E484K (B) to the ACE2 number of interfacing residues between hACE2 and S protein when applied to the reference sequence (grey), in tandem with variant defining groups of mutations (black). Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues.

The mutation A701V, like E484K appears in two variants, the VOI B.1.526 and the VOC B.1.351 (Fig.8). B.1.526 presents the lowest ACE2 interfacing residues off all variants tested, while B.1.351 the highest. When isolated, this mutation presents an identical result to B.1.351 in terms of the number of interfacing residues, however, the residues themselves vary.

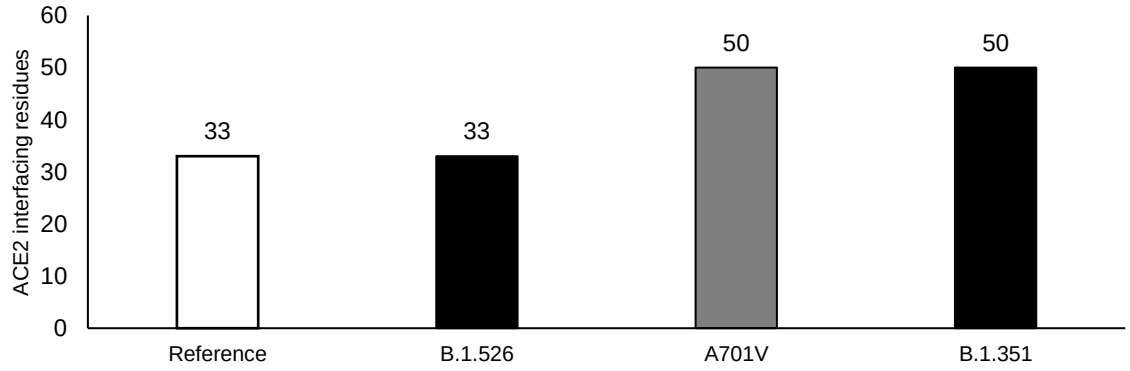


Fig.8 – Effect of the point mutation A701V to the ACE2 number of interfacing residues between hACE2 and S protein when applied to the reference sequence (gray), in tandem with variant defining groups of mutations (black). Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues.

The effect of point mutations greatly varies. This aspect is more visible outside of the RBD (labelled in grey in Fig.9), where a discrepancy of 16 ACE2 interfacing residues is observed between W152C and A701V. All mutations tested, that are present in known variants of SARS-CoV-2, presented a higher number of ACE2 interfacing residues than the wild type.

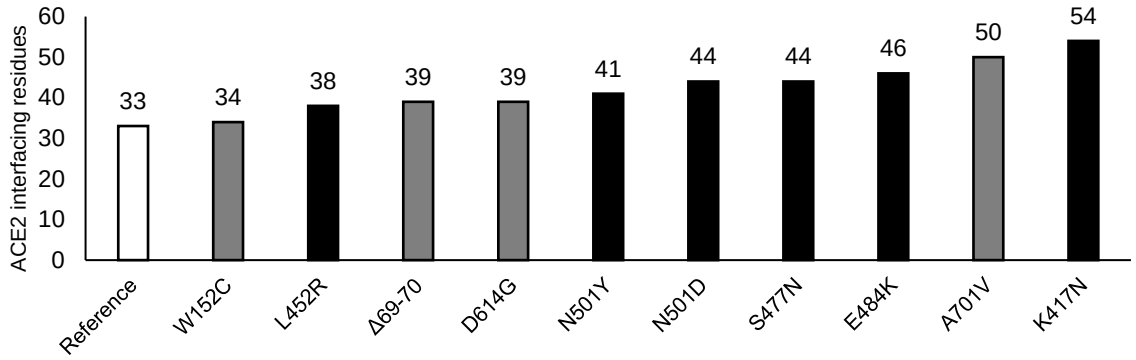


Fig.9 – Effect of the point mutations to the ACE2 number of interfacing residues between hACE2 and S protein when applied to the reference sequence. Mutations in RBM are represented in black and mutations outside the RBM are represented in grey. Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues.

4.4 SARS-CoV-2 variants and point mutation in core residues

Considering the outputs of PDBePISA for all variants and point mutations of SARS-CoV-2's S protein, a total of 19 S protein and 14 ACE2 core residues were found (interfacing frequency ≥ 90%). It is of note that the S protein presented three times more sites that always interfaced than ACE2, while the majority of ACE2 core residues don't present a frequency of 100% (Table 2).

Table 2 – List of core residues for both hACE2 and the S protein with frequencies equal or above 90% and 100%. For the S protein, the position number is based on the reference SARS-CoV-2 protein sequence.

Protein	≥90% frequency	100% frequency
hACE2	31, 37, 42, 324, 325, 354, 355, 386	30, 34, 35, 38, 41, 353
S protein	403, 484, 486	417, 449, 453, 455, 456, 487, 489, 490, 493, 494, 495, 496, 498, 501, 505

4.5 SARS-CoV-2 predictive model for infection

Thirteen species were confirmed to be susceptible to SARS-CoV-2 at the time of this study. The results viewed in PDBePISA of the interface of these species' ACE2 with SARS-CoV-2 S protein, revealed 13 residues of ACE2 interfacing in ≥ 90% of the species, 10 of which in 100% (Table 3). The same results yielded 25 SARS-CoV-2 S protein residues that always interface when docked with ACE2 of species confirmed to be infected by the virus. Additionally, two other residues presented a percentage of interface over 90%.

Table 3 – List of core residues for both SARS-CoV-2 's S protein and ACE2 of the positive control. The position number is based on the reference hACE2 protein sequence.

Protein	≥90% frequency	100% frequency
ACE2	30, 42, 393	31, 33, 34, 35, 37, 38, 41, 353, 354, 386
SARS-CoV-2 S protein	453, 486	403, 405, 408, 417, 446, 449, 455, 456, 473, 475, 483, 484, 485, 486, 487, 488, 489, 493, 494, 495, 496, 498, 500, 501, 505

The average number of interfacing residues, for susceptible species, when considering both proteins is 83.8 with a standard deviation of 7.2. This total is composed of a mean of 41.2 and 42.6 interfacing residues for S and ACE2 proteins, respectively. *Rousettus aegyptiacus* is the species with the highest S protein interfacing residues (45), whereas, *Macaca fascicularis* has the highest number for ACE2 and overall interface, 48 and 91, respectively (Fig.10). Of the confirmed infected species, *Homo sapiens* presents the lowest total and ACE2 interfacing residues, 71 and 33, respectively.

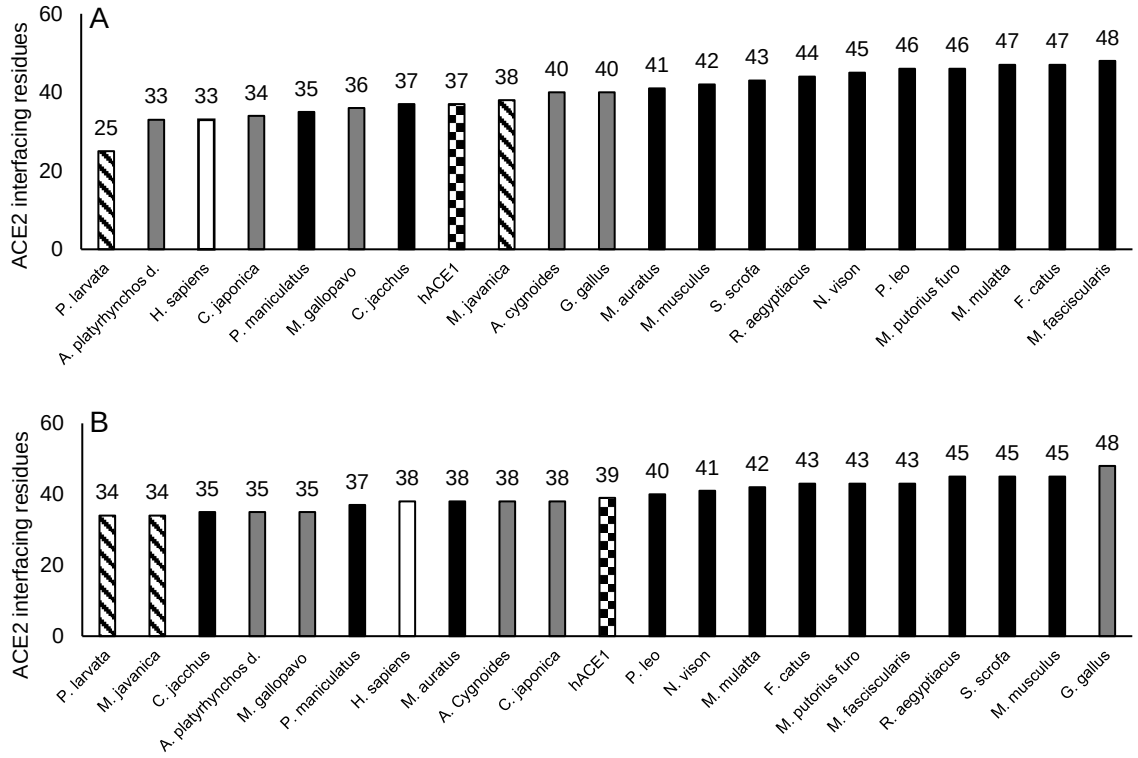


Fig.10 – ACE2 (A) and S protein (B) interfacing residues between SARS-CoV-2 S protein and Homo sapiens (white), susceptible species (black), non-susceptible species (grey), species with non-documented infection (stripes) and with hACE1 (squares). Data extracted from PDBePISA of the HADDOCK structure with the greatest number of S protein interfacing residues.

In comparison, in the list of species proven to not be susceptible by SARS-CoV--2, the total number of interfacing residues is 75.4, with a standard deviation of 7.1. For the S protein, the mean is 38.8, while for the ACE2, that number is 36.6, both lower than in the susceptible species. Statistical differences using the non-parametric Mann-Whitney U test (*P*=0.019) were only found for the number of ACE2 interfacing residues (Fig.11). Susceptible species present a number of ACE2 interfacing residues 1,16 times higher than non-susceptible species.

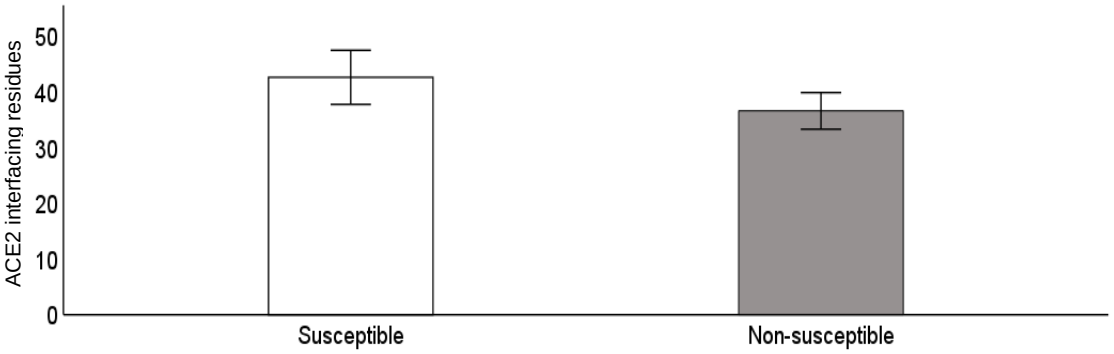


Fig.11 – Mean number of ACE2 interfacing residues with SARS-CoV-2's S protein for susceptible (white) and non-susceptible species (grey). Error bars represent one standard deviation. Data processed in SPSS.

4.5.1 SARS-CoV-2 – ACE2 core residues

The intersection of Tables 2 and 3 gives the true list of the SARS-CoV-2 – ACE2 core residues. For the variants', non-identical, but functional S proteins, the interfacing residues with high frequency, correspond to those that are needed for the protein docking. Similarly, the same conclusion can be made by using several ACE2 of species susceptible to SARS-CoV-2, i.e., documented cases of different ACE2-S protein docking. Hence, the SARS-CoV-2 variant and susceptible species sample, yielded 11 ACE2 and 16 S protein core residues (Table 4).

Table 4 – List of the intersected core residues ACE2 and S protein found for SARS-CoV-2 variants and SARS-CoV-2 susceptible species.

Protein	Core residues
ACE2	30, 31, 34, 35, 37, 38, 41, 42, 353, 354, 386
SARS-CoV-2 S protein	403, 417, 449, 455, 456, 484, 486, 487, 489, 493, 494, 495, 496, 498, 501, 505

It is of note that, of the 11 ACE2 residues, six (30, 34, 42, 353, 354 and 386) aren't conserved in the consensus sequence made with all sequences that form the ACE2 cluster previously described, while only three (30, 34 and 38) show this behaviour in mammals. For the S protein, relevant SARS-CoV-2 variants (VOCs and VOIs) are documented to have mutations in residues 417, 484 and 501 of their S proteins (Table 1).

4.5.2 Large scale predictive model

The positions 30, 34 and 354 presented interfacing residues with very diversified characteristics. Position 30 presented negatively charged polar, as well as non-polar hydrophobic small amino acids. For positions 34 and 354, aromatic, charged polar, aliphatic or hydrophobic residues could appear as interfacing. Hence, from the 11 ACE2 core residues, only 8 were used to assess candidates of ACE2 recognition from SARS-CoV-2 (Table 5). Of the remaining eight, 386 was the only to be treated as conserved. All sequences with alanine would have this site as interfacing, while species with serine (found in the negative control) would not. In contrast, site 31 was the most variable, presenting various types of polar amino acids. Apart from 386, all the sites considered present as interfacing polar amino acids.

Table 5 – List of interfacing amino acids for the core residues in ACE2 and their respective common characteristic. Lack of a common characteristic signifies a very diverse site, with the exception of site 386, which is a conserved one.

Residue site	Interfacing amino acid	Characteristic
30	[ADEN]	-
31	[EKNT]	Polar
34	[HLQSTVY]	-
35	[ER]	Big hydrophilic charged polar
37	[EQ]	Big hydrophilic polar
38	[DE]	Negatively charged polar
41	[HY]	Aromatic polar
42	[EQ]	Big hydrophilic polar
353	[HK]	Hydrophobic positively charged polar
354	[DGHNQR]	-
386	A	-

The attention turned to these eight sites of the ACE2 sequences obtained from NCBI in *Sequence retrieval*. To the 295 sequences, the *Panthera leo* and *Neovison vison* (confirmed infected species) sequences from *Ensembl* were added. A total of 109 species (37.1%) matched the amino acid characteristic pattern described in Table 5 for the eight sites analysed. Of these, 102 are mammals, six birds, all Passeriformes, and one reptile. Primates are the most prevalent with 25 species, followed by carnivores with 21 species from felines to canines and mustelids. Other prevalent orders include cetaceans and artiodactyls, both with 10 species. Additionally, five bat species, a hypothesised origin for SARS-CoV-2, also compose this candidates list.

Damas et al. (2020) used 25 known contact residues between hACE2 and SARS-CoV-2 S protein and scored a species susceptibility to infection based on the similarity of those residues to hACE2. The core residues here determined, apart from site 386, are present in the list used by Damas et al. (2020). The score system was based on five categories: very low, low, medium, high and very high. Of the 410 species tested, and considering medium as a cut-off, their prediction resulted in 104 species (25.4%) with at least a moderate likelihood of being susceptible to the virus, all of which mammals. Comparing the results of this method with the ones gathered by Damas et al. (2020), from the 109 species deemed as candidates for SARS-CoV-2 recognition, 9 were not contemplated in Damas et al. (2020)'s sample, 35 scored very low, 20 low, 24 medium, 12 high and 16 very high (Supplementary Table 1). Again, considering medium score as a cut-off value, both results agree in 55% of the cases. Interestingly, Damas' results show that all bats and pangolins (also a hypothesised SARS-CoV-2 origin) in their sample scored low or very low. Furthermore, all non-mammals in the sample scored very low.

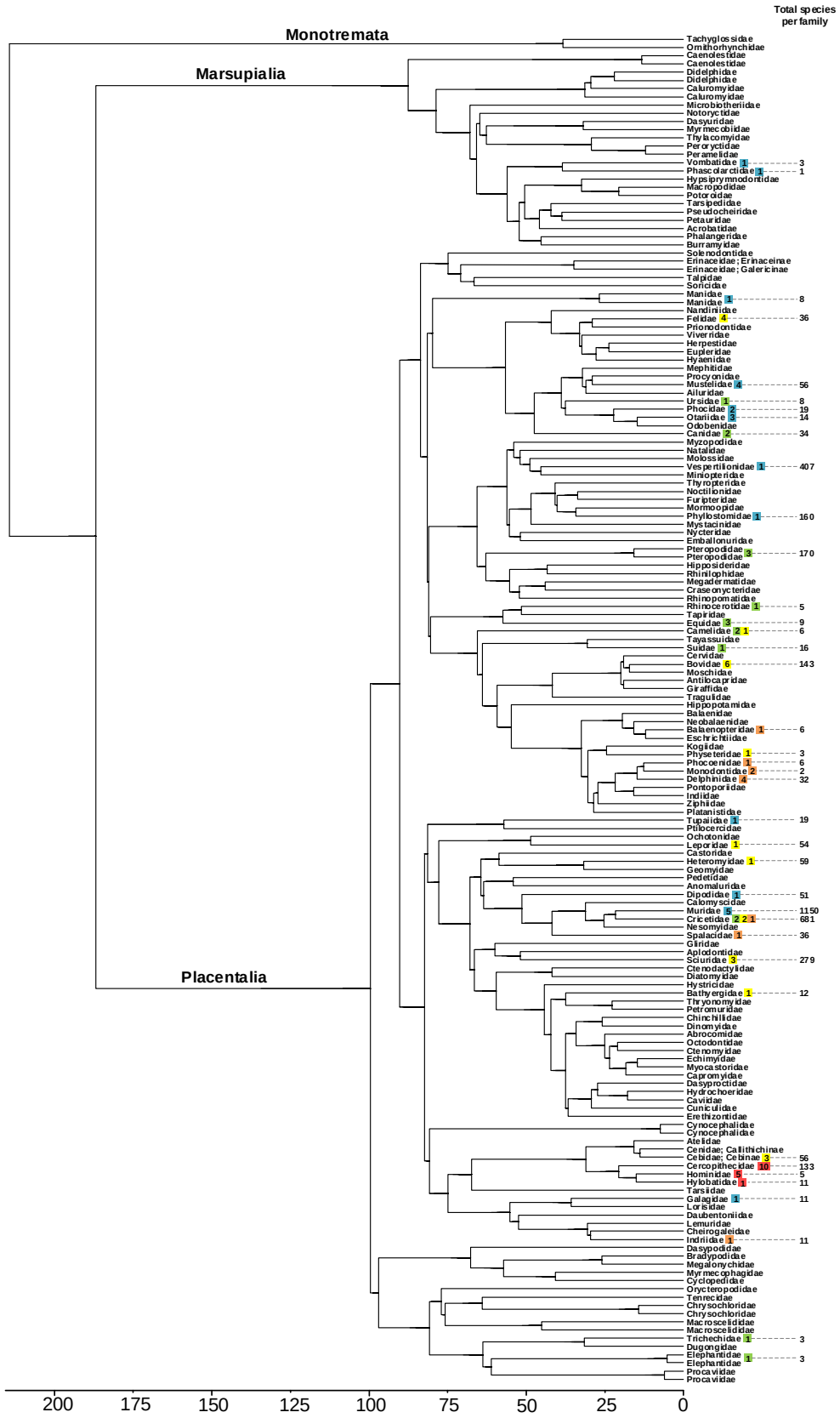


Fig.12 – Phylogenetic time tree of mammalian families adapted from Meredith et al. (2011). The number of candidates for SARS-CoV-2 – ACE2 recognition for a given family is displayed in a coloured square. The colour code correspond to the Damas et al. (2020) attributed score: blue – Very low; green – Low; yellow – Medium; orange – High; red – High. The number of species recognised for a given family with found candidates is also displayed (Myers et al., 2021).

Consulting Fig.12, it is clear that the found candidates for potential ACE2 recognition by SARS-CoV-2 are well dispersed. They were detected in Marsupialia and various clusters of Placentalia, although absent in the order Monotremata. Only the families Camelidae and Cricetidae possess within them candidates with different Damas et al. (2020) scores.

4.6 *Coronaviridae*

4.6.1 Identifying similar SARS-CoV-2's S protein in *Coronaviridae*

The S protein sequences downloaded from NCBI "Assembly" and ViPR, processed in SEDA, were used to produce Neighbor-joining trees in MEGAX. Both trees presented long branch lengths with low bootstrap values, making the identification of the most phylogenetically close S proteins unfeasible. Hence, the application of the *blastp* operation of SEDA to the biggest dataset, ViPR's. The 10 sequences with the highest score were retrieved. Of the 10 different *Coronaviridae* species obtained, three were able to use ACE2 as their cell receptor (RaTG13, YN2018B and SARS-CoV TWS) and seven were expected to not recognise this protein.

By analysing the residues that always interface in SARS-CoV-2 variants and those that only establish a bond with wild type, it is clear that none of the sequences retrieved had all these positions involved in the interface. Furthermore, species that were uncapable of binding to ACE2 had alignment gaps in three analysed positions of their S protein, 483, 484 and 486, the last two interfacing residues for all SARS-CoV-2 variants with hACE2. It is of note that the highest variation in interfacing residues compared to wild type appears in the S protein.

Viewing the alignment of the S proteins of these CoVs species in the positions considered, sites 489Y and 496G are conserved in species that are able to use ACE2. Additionally, positions 456H, 493S, 494T and 501V are conserved in species that don't use ACE2. This conservation seems to have no link with their ability to become interfacing residues, however, it might be related to the ability to bind to ACE2.

4.6.2 Use of ACE2

I) <u>Pattern finding</u>

Restraining the alignment of the previously obtained 10 *Coronaviridae* species to the S protein's RBM (positions 438-506 of SARS-CoV-2 wild type (Lan et al., 2020)), sites 450N, 468I, 488C, 491P, 500T, 502G and 504G join 489Y and 496G as conserved

position in species that are able to bind to ACE2. Also, the positions 439A, 440K and 459T, together with 456H, 493S, 494T and 501V, correspond to conserved positions of species unable to bind to ACE2.

Since variable alignment gaps occur until position 486, the region for pattern finding able to characterise a given species on their binding capabilities to ACE2 is limited to the region 488-506. Using the list of species given by Wells et al. (2020), the shortest pattern found to be able to separate the species of *Sarbecovirus* proven to be able to bind to ACE2 was CYX(6)GX(3)T[^V]. This pattern was not found in 25 species unable to bind to ACE2, while being present in all 13 species that were. Additionally, the pattern supported the clusters found by the same authors when using just the RBM region of the S protein (Supplementary Table 2).

Doing a *PHI-blast* in NCBI, using the SARS-CoV-2 S protein as query, and the pattern described here, yielded 139 matches, from the hosts *Homo sapiens*, *Manis javanica*, *Mus musculus, Paguma larvata*, *Rhinolophus affinis, Rhinolophus sinicus* and *Sus scrof*a. Similarly, searching the pattern in the ViPR dataset led to the retrieval of 92 matches. While some sequences were redundant within and across both datasets, different isolates from the same species or synthetic recombinants, no evidence of an entry unable to bind to ACE2 was found.

II) Structural similarities

In addition to the presence of the pattern described above, on average, species that are able to recognise and bind to the ACE2 protein, present a higher TM-score when pairing their RBM to SARS-CoV-2's, compared to those that don't (Supplementary Table 3). Normalized for the length of SARS-CoV-2's S protein and using the non-parametric Mann-Whitney U test, the average TM-score of the RBM of species able to bind to ACE2 is 0.904, a value statistically different (P<0.001) from 0.614, the average of the ones who cannot (Fig.13A). It is of note that overlapping of TM-scores was observed in the sample, such as the case of BM48-31 that presents a TM-score higher than those of WIV1 and RaTG13 (Fig.13B).

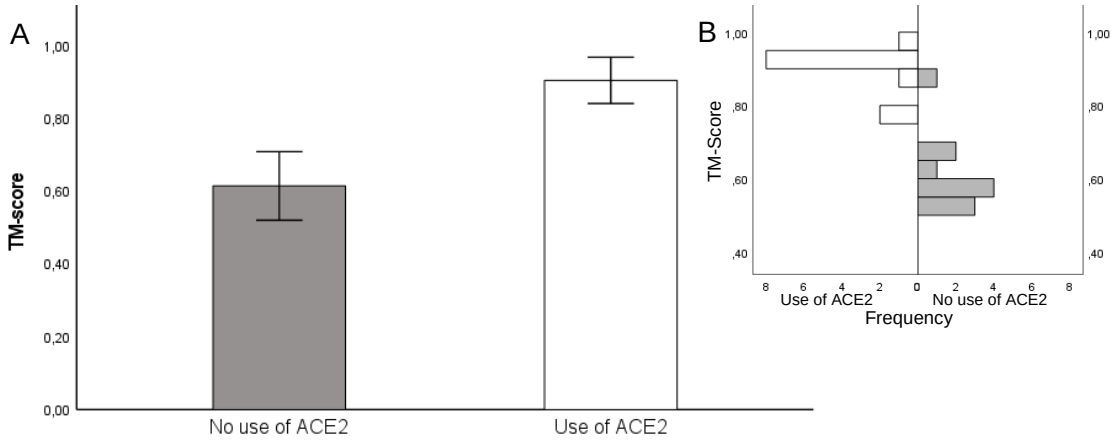


Fig.13 – Average TM-score for the RBM of species of coronavirus that are able (white) or unable (grey) to use the ACE2 protein to host cell entry, when paired with SARS-CoV-2. The score is normalized for the length of SARS-CoV-2's RBM. Error bars represent one standard deviation. Data processed in SPSS.

4.7 Positively selected sites

The approach used to determine the sites under positive selection worked better when employing FUBAR and codeML, rather than OMEGAMAP. More specifically, with the exception of a couple of cases, OMEGAMAP, produced large sequences of consecutive sites as output, signalling a region instead of a particular position. For SARS-CoV-2 100 random samples, FUBAR and codeML weren't always able to produce positively selected sites, the samples where this occurred were still considered valid. In total, FUBAR produced a total of 111 different sites under positive selection for SARS-CoV-2 (Table 6). Of these, out of the 100 random samples, 99 had a frequency of less than 25%, eight between 25 and 75% and four with a frequency over 75%. On the other hand, codeML, identified more sites, 182, however, 179 possessed a frequency lower than 25% and three with frequencies not exceeding 75% (Supplementary Table 4). Using codeML, no site with a frequency above 75% for our sample was found (Table 6).

Table 6 – Number of positively selected sites in SARS-CoV-2 and their respective frequency found using the codeML and FUBAR methods.

Method	Frequency		
	<25%	25%≤X<75%	≥75%
codeML	179	3	-
FUBAR	99	8	4

Despite the difference in frequency observed when using FUBAR or codeML, of the 12 sites with higher frequencies obtained with the former, eight are present in the top 12 of the latter. Considering both methods simultaneously, this approach was able to identify as positively selected all mutations reported for VOCs and VOIs present in Table 1. Additionally, the sites presented in Table 7, except site 732 belong to VOCs and/or VOIs. Mutation in the position 494 has been reported only after the data acquisition. These results further strengthen the findings.

Table 7 – Positively selected sites of SARS-CoV-2 with frequencies above 25% using codeML and FUBAR methods.

Method	Frequency	
	25%<X<75%	>75%
codeML	5, 681, 677	-
FUBAR	95, 732, 494, 138, 18, 26, 477, 681	5, 484, 501, 677

5. Discussion

5.1 Transmissibility rates of SARS-CoV-2 variants

Data regarding transmissibility rates of SARS-CoV-2 variants is somewhat limited. Measures such as quarantine and vaccination impaired its real time observation and quantification. Specifically regarding VOIs, the transmissibility rate was regarded here as equal to the wild type. This decision was based on the fact that a lower transmissibility rate, i.e., a lower fitness would contribute for the variant's disappearance. It is also important to note that a given variant may be treated as a VOI by a number of factors of which higher transmissibility is one, but not a mandatory criterion. A particular point in the description of VOIs is the ability to effect therapeutics or immune escape, a characteristic further enhanced by VOCs. Another significant information is the ability to produce higher viral loads than other variants, such as the case of variant B.1.617.2 that presents a viral load 1000x higher than the ones observed in the initial outbreak in China (Li et al., 2021). Thus, a variant that is able to circumvent the vaccination or replicate faster may become more prevalent in a given region, increasing the transmissibility rate, even though it has nothing to do with the S protein – ACE2 interface/affinity.

The results regarding the core residues for ACE2 – S protein docking suggest that there is little variation regarding the docking of the S protein and ACE2 protein, given the large number of residues interfacing in 100% of the variants tested. That means that SARS-CoV-2´s S proteins with point mutations or groups of observed mutations aren't able to produce much variability in the set of the S protein interfacing residues, especially when compared to ACE2. The correlation between the number of ACE2 interfacing

residues and the transmissibility rate of SARS-CoV-2 variants described above, further supports the higher role of the ACE2 residues.

The S protein mutations produce, however, an effect in ACE2's interfacing residues. Although with no visible impact in the number and position of HB and SB, all point mutations tested (belonging to VOCs and/or VOIs) induced more ACE2 interfacing residues than the wild type. This increase varied from 3 to 63.4%, with the two highest values belonging to mutations inside and outside the RBM.

Of the VOIs and VOCs tested, none produced less ACE2 interfacing residues than the wild type. Given the already mentioned correlation regarding ACE2 residues, it supports the hypothesis of the adaptability of the virus with every variation that manages to implement itself and spread in a given population. Furthermore, the discrepancy of results between point mutations and variation of ACE2 interfacing residues shows that it is not enough to evaluate the impact of the point mutations. A case made clear by mutations E484K and A701V. Individually, they produce a considerable increase in ACE2 interfacing residues, however, both are present in the only variant with the same number of interfacing residues as the wild type. Labelling the impact of a single mutation purely based in their location, i.e., inside or outside the RBM is also erroneous. Therefore, for any given future mutation to SARS-CoV-2's S protein, the docking protocol here described should be applied, in order to more correctly assess the impact of said mutations.

5.2 ACE2 usage

Looking at the data regarding the 10 *Coronaviridae* species with the S proteins closest to SARS-CoV-2 (sequence wise), it would be logical to infer that none could infect humans. This statement is supported by no species establishing interface in all the S protein and hACE2 core residues. However, one entry of the sample is from SARS-CoV, a virus that infects humans and uses ACE2 for cell entry. Even though it had the most interfacing core residues, this implies that applying this criterion in an interspecific fashion is erroneous.

A pattern for ACE2 usage recognition was described by Carvalho and Alves (2020), the hydrophobic sequence YGFY. However, this sequence is not present in SARS-CoV-2, where the motif changes to YGFQ. In the same manner, the *Bat SARS-like CoV Rs7327,* that uses ACE2 also changes the motif to YGFF, identical to the one presented by *Bat CoV BM48-31*, a species that is documented as unable to recognise ACE2 (Hu et al., 2017). These cases strengthen the motif YGF while weakening the initial four amino acid one, given how species that are able to recognise ACE2 would be looked over by not presenting YGFY. Contrasting to this, from the unfiltered 1397 S

proteins sequences extracted from ViPR, the motif YGF appears 1568 times, more than once per sequence.

The pattern here suggested is conserved in all species tested and was not present in non-ACE2 binding species. Furthermore, given that it is large, in the same 1397 sequence dataset, it was only found once per S protein sequence, making it a better pattern to be used.

It is of note that the endemic hCoV-NL63 also uses ACE2 to infect host cells. However, unlike SARS-CoV and SARS-CoV-2, who are β-CoVs, this α-CoV interacts with ACE2 in a different way. Instead of one large domain, NL63 possesses different RBD in its S protein (Wu et al., 2009). Unlike all the ACE2 binding species tested, NL63 does not present the pattern found. The lack of sequence similarity and different binding method may justify this result. NL63 also binds less efficiently to ACE2, however, this result only gives information in a binary form and it does not provide evidence of the use of all ACE2s as receptor for cell entry. Ultimately, these findings support that, for the sample tested, all β-CoVs with the documented CYX(6)YX(3)T[^V] in their S proteins can bind to ACE2.

Furthermore, analysing the structural similarity to the RBM of SARS-CoV-2 of different CoVs, a clear difference can be seen regarding their ability to bind to ACE2. It was found that species able to bind to ACE2 have a RBM with a 3D structure more closely resembling SARS-CoV-2, than those who aren't. Again, *Bat CoV BM48-31* shows a contrary position to the norm, having a TM-score surpassing two species able to bind to ACE2. One of said species is RaTG13, which, genome wise, closely resembles SARS-CoV-2. This further exemplifies the difficulty of excluding species such as BM48-31.

So, it is to be expected that a species that is able to bind to ACE2 presents in their S protein the pattern CYX(6)YX(3)T[^V], while also having a high similarity, structure wise, to SARS-CoV-2's RBM.

5.3 Large scale susceptibility model

The application of the docking protocol to susceptible and non-susceptible species of SARS-CoV-2 allowed a view where the ACE2 protein was the main variable. The ACE2 interfacing residues, believed to have a role in the transmissibility of SARS-CoV-2 variants, have a larger variability in susceptible species and is statistically higher than in non-susceptible species. Interestingly, one of the smallest numbers belongs to the docking with hACE2.

Nonetheless, the *Sus scrofa* case reveals that some species can be infected *in vitro*, however, *in vivo* they can be seen as immune. Thus, it cannot be discarded that other species can behave in the same manner, being classified as non-susceptible by

only doing *in vivo* studies (Suarez et al., 2020). If this is the case, and some information can be obtained by the number of ACE2 interfacing residues, then the low number observed in the human case was to be expected since the reference SARS-CoV-2 had no time to adapt to the new host. It would also mean that all the other species would have more suitable conditions (protein – protein interaction wise) to harbour the virus. This ability would be impaired by other factors such as the immune system of the host. Additionally, specific sites in the Spike and ACE2 proteins establish an interface in susceptible species and are absent in non-susceptible ones. In particular, hACE2's residue 386, with interaction absent in the orthologs of the non-susceptible species and hACE1. In fact, this residue is very conserved and only establishes interface when alanine is present in that position.

For the large scale approach to determine the ability of SARS-CoV-2 to recognise a species' ACE2, the protein docking is sacrificed for speed, therefore only the ACE2, in particular, the determined core residues are considered. It is hypothesised that a site with a conserved amino acid will behave as interfacing in the protein docking, however, that may not always be the case. The same can be said for the S protein.

Like Damas et al. (2020), cetaceans, primates, bovines and felines are generally regarded as candidates for susceptibility. In contrast, the method here described considered as candidates mustelids, equids, rodents, in particular murids and otarrids. The average score attributed by Damas to these groups was very low. *Mustela putorius furo* being a confirmed SARS-CoV-2 infected species and the Denmark mink massacre show mustelids as a concern group, corroborating the results here shown. Additionally, seven non-mammals were also scored positively here, six bids and one reptile. Notably, by applying the protein protocol, the reptile *Zootoca vivipara*, who was not present in Damas et al. (2020) dataset, as well as *Ficedula albicolis*, *Serinus canaria* and *Camarhynchus parvulus* (birds) show all core residues of both ACE2 and SARS-CoV-2's S protein as interfacing.

As stated before, contrary to Damas et al. (2020), this study's results show a stronger support to the origin of SARS-CoV-2 in bats, as *Eptesicus fuscus*, *Phyllostomus discolor*, *Pteropus alecto*, *Pteropus vampyrus* and *Rousettus aegyptiacus* are deemed as candidates for ACE2 recognition by SARS-CoV-2. However, these results are to be taken as preliminary, since by applying the protein docking protocol to *Manis javanica*, a candidate for susceptibility, residues 41 and 42 of ACE2 don't appear as interfacing. Similarly, the birds *Cyanistes caeruleus, Sturnus vulgaris* and *Zonotrichia albicolis* don't present the ACE2 core residue 386 as interfacing, a result further supported by the lack of interfacing residues around that area. Therefore, signalling the importance of applying the whole protocol. Notably, only ACE2 core residues were regarded as non-interfacing for non-mammal species. This result further supports the importance of ACE2 over the S protein for SARS-CoV-2 recognition and the conserved way that the S protein docks

to ACE2. The same protocol was made using the ACE1 sequences extracted in the ACE2 sequence retrieval. No candidate for infection was found, therefore excluding this protein as a possible candidate for viral recognition and subsequent cell entry.

The nature of this predictive model does not allow the definition provided by the scoring system of Damas et al. (2020), however it brings forward the possibility of SARS-CoV-2 infecting non-mammals, in particular a reptile and Passeriformes. It is also able to score groups like bats and mustelids in a manner more in tune with humanity's actions and concerns. Since β-CoVs are reported to only infect mammals, this result should be reviewed and confirmed to assess the validity of the model outside mammals, as described above.

The candidate mammal hosts appear very dispersedly in the mammalian phylogenetic tree. Furthermore, some families present more than one candidate species with different scores by Damas. Thus, inferring the likelihood of a given species being susceptible to SARS-CoV-2 due to phylogenetic proximity is unadvisable.

It is important to note that this analysis had some limitations. The number of confirmed susceptible and non-susceptible species was very low, weakening the statistical analysis conducted. Plus, only the initial SARS-CoV-2 S protein was tested, hence, the impact of the newly emerged variants is unknown.

5.4 Positively selected sites

Almost all PSS that appeared with a frequency above 25% for SARS-CoV-2 determined using the IPSSA pipeline belong to VOCs and VOIs. This supports the idea that these mutations increase in frequency due to positive selection. In the case of sites 494 and 732, they should be studied to infer their impact in case of mutations in present or future variants. The mutation S494P has recently been detected in some samples of the variant B.1.1.7, supporting the prediction power of this method. For site 732, however, no significant mention has been found, despite this, if the trend is followed, it is expected to appear in a future variant. The low frequency of sites like 614, considered the first major mutation, can be justified by its fixation in the populations. Mutations that are now at 100% frequency appear as PSS because all 100 random samples included sequences similar to the wild type (from 2019).

The analysis of the PSS obtained for the 50 samples belonging to other CoV species was not able to support and predict with high accuracy PSS discovered for SARS-CoV-2. For the sample tested, codeML attributes the highest number of identified PSS to the region 1-10 while FUBAR to the regions 21-30 and 481-490. These regions cover the sites 5, 26 and 484, PSS obtained using SARS-CoV-2 datasets. Although these results sound promising, a given PSS is only shared by a maximum of two species.

This doesn't allow resolution of data, and thus, it makes it difficult to infer and extrapolate data regarding SARS-CoV-2's PSS. A possible explanation could be the lack of data. Some species of CoV treated only had the minimum number of sequences for the pipeline to run. Others had sequences with considerable diversity, possibly due to recombination factors, poor annotation or just contamination in the sequencing process. It is not possible to ensure good results by having more data, nevertheless the dataset that yielded results consisted of a number of sequences several orders of magnitude higher than all the others.

Making a file with the largest sequence of each dataset, and including a SARS-CoV-2 sequence, yields no results. Since the sequences belonging to different species have such large diversity, the sequence has a large number of alignment gaps. Even with some alignment regions found, the huge diversity hinders the identification of PSS. This scenario would not chance depending on the number of sequences used, since gaps exist even intra species, therefore the interspecific approach for the detection of PSS was abandoned.

Thus, the assessment and prediction of sites that may characterize SARS-CoV-2 VOCs and VOIs can be made using SARS-CoV-2 sequence datasets. Some imprecise data can be inferred using datasets with only one species of CoV and no information can be retrieved using datasets with several different species.

6. Conclusion

The results obtained by this study show that the bioinformatics approach tested proved to be very effective in the detection of patterns. In fact, the broadening of the baseline methodology for the creation of the predictive models, allowed the exploration of the *Coronaviridae* from a variety of different angles. Namely, the exploration of the S protein of different CoVs resulted in the identification of those who are able to use the protein ACE2 as a cell entry mechanism. Indeed, the pattern CYX(6)YX(3)T[^V] in the S protein of a given CoV was never associated with the inability to use ACE2. Furthermore, by producing the 3D model of the S protein of different CoVs, it was possible to assess that the RBM's 3D structure closer to that of SARS-CoV-2 is likewise linked to CoV species that use ACE2.

Focusing specifically on SARS-CoV-2, a factor for assessing the transmissibility rate of SARS-CoV-2 variants was found to be the number of hACE2 interfacing residues, supported by a correlation in the sample tested. The mutations that characterize these variants could also be predicted by evaluating the sites under positive selection. Thus, the results here presented contribute not only to the prediction of threats of other CoVs, but also, the increased transmissibility rate of existing SARS-CoV-2 variants and the assessment of the next major mutations. Moreover, a list of candidate species for SARS-CoV-2 susceptibility more closely representative of reality was obtained. By including more bats and mustelids, one being a candidate for the origin of SARS-CoV-2 and the other associated with positive infection and transmission, this result represents a step forward from Damas et al. (2020) in regards to this field.

The ability to produce this diversity of data is a testament to the flexibility and strength of the methods here described. An approach that has its biggest limitation in the use of third party servers and their policies. However, it is of note that the acquisition of data was made without major delays and completely free of charge.

Undoubtedly, the promising results should be updated with newer SARS-CoV-2 variants, namely their respective transmissibility rates and set of mutations, giving more statistical support to the findings. Furthermore, it is also important to introduce more information regarding the confirmed host range of SARS-CoV-2.

A prospected task is to evaluate the changes in the host range of SARS-CoV-2 due to the mutations that occur in the S protein, possibly leading to the creation of different lists of susceptible species. An effort should also be made to further refine the results in order to incorporate threat levels instead of a binary result. Incorporating more recently available data, an expedite and inexpensive process using the methods in this study, should make these future milestones achievable.

An extended use of these methodologies to other viruses is also a possibility. Given its predictive nature, it could change the way that we respond or react to future difficulties to contain contagious diseases, such as those with aerial transmission.

7. References

Ahmad, T., Khan, M., Haroon, Musa, T. H., Nasir, S., Hui, J., Bonilla-Aldana, D. K., & Rodriguez-Morales, A. J. (2020). COVID-19: Zoonotic aspects. *Travel Medicine and Infectious Disease*, *36*. https://doi.org/10.1016/j.tmaid.2020.101607

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nature Medicine*, *26*(4), 450-452. https://doi.org/10.1038/s41591-020-0820-9

Anthony, S. J., Gilardi, K., Menachery, V. D., Goldstein, T., Ssebide, B., Mbabazi, R., Navarrete-Macias, I., Liang, E., Wells, H., Hicks, A., Petrosov, A., Byarugaba, D. K., Debbink, K., Dinnon, K. H., Scobey, T., Randell, S. H., Yount, B. L., Cranfield, M., Johnson, C. K., Baric, R. S., Lipkin, W. I., & Mazet, J. A. K. (2017). Further Evidence for Bats as the Evolutionary Source of Middle East Respiratory Syndrome Coronavirus. *mBio*, *8*(2). https://doi.org/10.1128/mBio.00373-17

Arbour, N., Côté, G., Lachance, C., Tardieu, M., Cashman, N. R., & Talbot, P. J. (1999). Acute and Persistent Infection of Human Neural Cell Lines by Human Coronavirus OC43. *Journal of Virology*, *73*(4), 33-38. https://doi.org/10.1128/JVI.73.4.3338-3350.1999

Azhar, E. I., El-Kafrawy, S. A., Farraj, S. A., Hassan, A. M., Al-Saeed, M. S., Hashem, A. M., & Madani, T. A. (2014). Evidence for Camel-to-Human Transmission of MERS Coronavirus. *New England Journal of Medicine*, *370*(26), 2499-2505. https://doi.org/10.1056/NEJMoa1401505

Berman, H., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Structural & Molecular Biology*, *10*(12), 980-980. https://doi.org/10.1038/nsb1203-980

Bosch, B. J., van der Zee, R., de Haan, C. A., & Rottier, P. J. (2003). The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. *J Virol*, *77*(16), 8801-8811. https://doi.org/10.1128/jvi.77.16.8801-8811.2003

Carvalho, P. P. D., & Alves, N. (2020). *Featuring ACE2 binding SARS-CoV and SARS-CoV-2 through a conserved evolutionary pattern of amino acid residues*.

Centers for Disease Control and Prevention. (2021, 05/05/2021). *SARS-CoV-2 Variant Classifications and Definitions* Retrieved 06/05/2021 from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html?fbclid=IwAR3YwLYIOMz431yqTo-ZyEiG8V0ruYv_XkrgOZzZUzC77nbROg11bUfm5Qg

Corman, V. M., Ithete, N. L., Richards, L. R., Schoeman, M. C., Preiser, W., Drosten, C., & Drexler, J. F. (2014). Rooting the Phylogenetic Tree of Middle East Respiratory

Syndrome Coronavirus by Characterization of a Conspecific Virus from an African Bat. *Journal of Virology*, *88*(19), 112-197. https://doi.org/10.1128/JVI.01498-14

Corman, V. M., Muth, D., Niemeyer, D., & Drosten, C. (2018). Chapter Eight - Hosts and Sources of Endemic Human Coronaviruses. In M. Kielian, T. C. Mettenleiter, & M. J. Roossinck (Eds.), *Advances in virus research* (Vol. 100, pp. 163-188). Academic Press. https://doi.org/10.1016/bs.aivir.2018.01.001

Crackower, M. A., Sarao, R., Oudit, G. Y., Yagil, C., Kozieradzki, I., Scanga, S. E., Oliveira-dos-Santos, A. J., da Costa, J., Zhang, L., Pei, Y., Scholey, J., Ferrario, C. M., Manoukian, A. S., Chappell, M. C., Backx, P. H., Yagil, Y., & Penninger, J. M. (2002). Angiotensin-converting enzyme 2 is an essential regulator of heart function. *Nature*, *417*(6891), 822-828. https://doi.org/10.1038/nature00786

Cui, J., Li, F., & Shi, Z.-L. (2019). Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*, *17*(3), 181-192. https://doi.org/10.1038/s41579-018-0118-9

da Silva, P. G., Mesquita, J. R., de São José Nascimento, M., & Ferreira, V. A. M. (2021). Viral, host and environmental factors that favor anthropozoonotic spillover of coronaviruses: An opinionated review, focusing on SARS-CoV, MERS-CoV and SARS-CoV-2. *Science of The Total Environment*, *750*. https://doi.org/10.1016/j.scitotenv.2020.141483

Damas, J., Hughes, G. M., Keough, K. C., Painter, C. A., Persky, N. S., Corbo, M., Hiller, M., Koepfli, K.-P., Pfenning, A. R., Zhao, H., Genereux, D. P., Swofford, R., Pollard, K. S., Ryder, O. A., Nweeia, M. T., Lindblad-Toh, K., Teeling, E. C., Karlsson, E. K., & Lewin, H. A. (2020). Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proceedings of the National Academy of Sciences*, *117*(36), 22311-22322. https://doi.org/10.1073/pnas.2010146117

de Groot, R. J., Baker, S. C., Baric, R., Enjuanes, L., Gorbalenya, A. E., Holmes, K. V., Perlman, S., Poon, L., Rottier, P. J. M., Talbot, P. J., Woo, P. C. Y., & Ziebuhr, J. (2012). Family *Coronaviridae*. In A. King, M. Adams, E. Cartens, & E. Lefkowitz (Eds.), *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses Academic Press* (pp. 806-828).

Deng, X., Garcia-Knight, M. A., Khalid, M. M., Servellita, V., Wang, C., Morris, M. K., Sotomayor-González, A., Glasner, D. R., Reyes, K. R., Gliwa, A. S., Reddy, N. P., Sanchez San Martin, C., Federman, S., Cheng, J., Balcerek, J., Taylor, J., Streithorst, J. A., Miller, S., Kumar, G. R., Sreekumar, B., Chen, P.-Y., Schulze-Gahmen, U., Taha, T. Y., Hayashi, J., Simoneau, C. R., McMahon, S., Lidsky, P. V., Xiao, Y., Hemarajata, P., Green, N. M., Espinosa, A., Kath, C., Haw, M., Bell, J., Hacker, J. K., Hanson, C., Wadford, D. A., Anaya, C., Ferguson, D., Lareau,

L. F., Frankino, P. A., Shivram, H., Wyman, S. K., Ott, M., Andino, R., & Chiu, C. Y. (2021). Transmission, infectivity, and antibody neutralization of an emerging SARS-CoV-2 variant in California carrying a L452R spike protein mutation. *medRxiv*. https://doi.org/10.1101/2021.03.07.21252647

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., Aprea, G., Aury, J.-M., Bento, P., Bernard, M., Bocs, S., Campa, C., Cenci, A., Combes, M.-C., Crouzillat, D., Da Silva, C., Daddiego, L., De Bellis, F., Dussert, S., Garsmeur, O., Gayraud, T., Guignon, V., Jahn, K., Jamilloux, V., Joët, T., Labadie, K., Lan, T., Leclercq, J., Lepelley, M., Leroy, T., Li, L.-T., Librado, P., Lopez, L., Muñoz, A., Noel, B., Pallavicini, A., Perrotta, G., Poncet, V., Pot, D., Priyono, Rigoreau, M., Rouard, M., Rozas, J., Tranchant-Dubreuil, C., VanBuren, R., Zhang, Q., Andrade, A. C., Argout, X., Bertrand, B., de Kochko, A., Graziosi, G., Henry, R. J., Jayarama, Ming, R., Nagai, C., Rounsley, S., Sankoff, D., Giuliano, G., Albert, V. A., Wincker, P., & Lashermes, P. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, *345*(6201). https://doi.org/10.1126/science.1255274

Dominguez, C., Boelens, R., & Bonvin, A. M. J. J. (2003). HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society*, *125*(7), 1731-1737. https://doi.org/10.1021/ja026939x

Donoghue, M., Hsieh, F., Baronas, E., Godbout, K., Gosselin, M., Stagliano, N., Donovan, M., Woolf, B., Robison, K., Jeyaseelan, R., Breitbart, R. E., & Acton, S. (2000). A novel angiotensin-converting enzyme-related carboxypeptidase (ACE2) converts angiotensin I to angiotensin 1-9. *Circ Res*, *87*(5), 1-9. https://doi.org/10.1161/01.res.87.5.e1

European Centre for Disease Prevention and Control. (2020a). *COVID-19 situation update worldwide, as of week 52 2020*. https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases

European Centre for Disease Prevention and Control. (2020b). Rapid increase of a SARS-CoV-2 variant with multiple spike protein mutations observed in the United Kingdom – 20 December 2020. *ECDC: Stockholm*.

Faria, N. R., Claro, I. M., Andrade, P. S., Franco, L. A. M., Aguiar, R. S., Coletti, T. M., Silva, C. A. M., Sales, F. C., Manuli, E. R., Candido, D., Gaburo, N., Camilo, C. C., Fraiji, N. A., Crispim, M. A. E., Carvalho, M. P. S. S., Rambaut, A., Loman, N., Pybus, O., & Sabino, E. C. (2021). *Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings*

Fehr, A. R., Channappanavar, R., & Perlman, S. (2017). Middle East Respiratory Syndrome: Emergence of a Pathogenic Human Coronavirus. *Annual Review of Medicine*, *68*(1), 387-399. https://doi.org/10.1146/annurev-med-051215-031152

Forni, D., Cagliani, R., Clerici, M., & Sironi, M. (2017). Molecular Evolution of Human Coronavirus Genomes. *Trends Microbiol*, *25*(1), 35-48. https://doi.org/10.1016/j.tim.2016.09.001

Fournier, D., Luft, F. C., Bader, M., Ganten, D., & Andrade-Navarro, M. A. (2012). Emergence and evolution of the renin-angiotensin-aldosterone system. *J Mol Med (Berl)*, *90*(5), 495-508. https://doi.org/10.1007/s00109-012-0894-z

Galloway, S., Paul, P., MacCannell, D., Johansson, M., Brooks, J., MacNeil, A., Slayton, R., Tong, S., Silk, B., Armstrong, G., Biggerstaff, M., & Dugan, V. (2021). Emergence of SARS-CoV-2 B.1.1.7 Lineage — United States, December 29, 2020–January 12, 2021. *MMWR. Morbidity and Mortality Weekly Report*, *70*. https://doi.org/10.15585/mmwr.mm7003e2

Ge, X.-Y., Wang, N., Zhang, W., Hu, B., Li, B., Zhang, Y.-Z., Zhou, J.-H., Luo, C.-M., Yang, X.-L., Wu, L.-J., Wang, B., Zhang, Y., Li, Z.-X., & Shi, Z.-L. (2016). Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft. *Virologica Sinica*, *31*(1), 31-40. https://doi.org/10.1007/s12250-016-3713-9

Gorbalenya, A. E., Baker, S. C., Baric, R. S., de Groot, R. J., Drosten, C., Gulyaeva, A. A., Haagmans, B. L., Lauber, C., Leontovich, A. M., Neuman, B. W., Penzar, D., Perlman, S., Poon, L. L. M., Samborskiy, D. V., Sidorov, I. A., Sola, I., Ziebuhr, J., & Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, *5*(4), 536-544. https://doi.org/10.1038/s41564-020-0695-z

Graham, M. S., Sudre, C. H., May, A., Antonelli, M., Murray, B., Varsavsky, T., Kläser, K., Canas, L. S., Molteni, E., Modat, M., Drew, D. A., Nguyen, L. H., Polidori, L., Selvachandran, S., Hu, C., Pujol, J. C., Hammers, A., Chan, A. T., Wolf, J., Spector, T. D., Steves, C. J., & Ourselin, S. (2021). The effect of SARS-CoV-2 variant B.1.1.7 on symptomatology, re-infection and transmissibility. *medRxiv*. https://doi.org/10.1101/2021.01.28.21250680

Gryseels, S., De Bruyn, L., Gyselings, R., Calvignac-Spencer, S., Leendertz, F. H., & Leirs, H. (2020). Risk of human-to-wildlife transmission of SARS-CoV-2. *Mammal Review*. https://doi.org/10.1111/mam.12225

Gryseels, S., De Bruyn, L., Gyselings, R., Calvignac-Spencer, S., Leendertz, F. H., & Leirs, H. (2021). Risk of human-to-wildlife transmission of SARS-CoV-2. *Mammal Review*, *51*(2), 272-292. https://doi.org/10.1111/mam.12225

Gu, H., Chen, Q., Yang, G., He, L., Fan, H., Deng, Y.-Q., Wang, Y., Teng, Y., Zhao, Z., Cui, Y., Li, Y., Li, X.-F., Li, J., Zhang, N., Yang, X., Chen, S., Zhao, G., Wang, X., Luo, D., Wang, H., Yang, X., Li, Y., Han, G., He, Y., Zhou, X., Geng, S., Sheng, X., Jiang, S., Sun, S., Qin, C.-F., & Zhou, Y. (2020). Rapid adaptation of SARS-CoV-2 in BALB/c mice: Novel mouse model for vaccine efficacy. *bioRxiv*. https://doi.org/10.1101/2020.05.02.073411

Guan, Y., Zheng, B. J., He, Y. Q., Liu, X. L., Zhuang, Z. X., Cheung, C. L., Luo, S. W., Li, P. H., Zhang, L. J., Guan, Y. J., Butt, K. M., Wong, K. L., Chan, K. W., Lim, W., Shortridge, K. F., Yuen, K. Y., Peiris, J. S. M., & Poon, L. L. M. (2003). Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China. *Science*, *302*(5643), 276. https://doi.org/10.1126/science.1087139

Hamre, D., & Procknow, J. J. (1966). A New Virus Isolated from the Human Respiratory Tract. *Proceedings of the Society for Experimental Biology and Medicine*, *121*(1), 190-193. https://doi.org/10.3181/00379727-121-30734

Heald-Sargent, T., & Gallagher, T. (2012). Ready, set, fuse! The coronavirus spike protein and acquisition of fusion competence. *Viruses*, *4*(4), 557-580. https://doi.org/10.3390/v4040557

Heegaard, P. M. H., Sturek, M., Alloosh, M., & Belsham, G. J. (2020). Animal Models for COVID-19: More to the Picture Than ACE2, Rodents, Ferrets, and Non-human Primates. A Case for Porcine Respiratory Coronavirus and the Obese Ossabaw Pig. *Frontiers in Microbiology*, *11*, 23-33. https://doi.org/10.3389/fmicb.2020.573756

Hoffmann, M., Kleine-Weber, H., Krüger, N., Müller, M., Drosten, C., & Pöhlmann, S. (2020). The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. *bioRxiv*. https://doi.org/10.1101/2020.01.31.929042

Hu, B., Zeng, L.-P., Yang, X.-L., Ge, X.-Y., Zhang, W., Li, B., Xie, J.-Z., Shen, X.-R., Zhang, Y.-Z., Wang, N., Luo, D.-S., Zheng, X.-S., Wang, M.-N., Daszak, P., Wang, L.-F., Cui, J., & Shi, Z.-L. (2017). Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLOS Pathogens*, *13*(11). https://doi.org/10.1371/journal.ppat.1006698

Isabel, S., Graña-Miraglia, L., Gutierrez, J. M., Bundalovic-Torma, C., Groves, H. E., Isabel, M. R., Eshaghi, A., Patel, S. N., Gubbay, J. B., Poutanen, T., Guttman, D. S., & Poutanen, S. M. (2020). Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. *bioRxiv*. https://doi.org/10.1101/2020.06.08.140459

Kirchdoerfer, R. N., Cottrell, C. A., Wang, N., Pallesen, J., Yassine, H. M., Turner, H. L., Corbett, K. S., Graham, B. S., McLellan, J. S., & Ward, A. B. (2016). Pre-fusion structure of a human coronavirus spike protein. *Nature*, *531*(7592), 118-121. https://doi.org/10.1038/nature17200

Kirchdoerfer, R. N., Wang, N., Pallesen, J., Wrapp, D., Turner, H. L., Cottrell, C. A., Corbett, K. S., Graham, B. S., McLellan, J. S., & Ward, A. B. (2018). Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-34171-7

Kowalczuk, S., Bröer, A., Tietze, N., Vanslambrouck, J. M., Rasko, J. E. J., & Bröer, S. (2008). A protein complex in the brush-border membrane explains a Hartnup disorder allele. *The FASEB Journal*, *22*(8), 2880-2887. https://doi.org/10.1096/fj.08-107300

Lam, T. T.-Y., Jia, N., Zhang, Y.-W., Shum, M. H.-H., Jiang, J.-F., Zhu, H.-C., Tong, Y.-G., Shi, Y.-X., Ni, X.-B., Liao, Y.-S., Li, W.-J., Jiang, B.-G., Wei, W., Yuan, T.-T., Zheng, K., Cui, X.-M., Li, J., Pei, G.-Q., Qiang, X., Cheung, W. Y.-M., Li, L.-F., Sun, F.-F., Qin, S., Huang, J.-C., Leung, G. M., Holmes, E. C., Hu, Y.-L., Guan, Y., & Cao, W.-C. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*, *583*(7815), 282-285. https://doi.org/10.1038/s41586-020-2169-0

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

Lau, S. K., Li, K. S., Tsang, A. K., Lam, C. S., Ahmed, S., Chen, H., Chan, K. H., Woo, P. C., & Yuen, K. Y. (2013). Genetic characterization of Betacoronavirus lineage C viruses in bats reveals marked sequence divergence in the spike protein of pipistrellus bat coronavirus HKU5 in Japanese pipistrelle: implications for the origin of the novel Middle East respiratory syndrome coronavirus. *J Virol*, *87*(15), 8638-8650. https://doi.org/10.1128/jvi.01055-13

Lau, S. K. P., Woo, P. C. Y., Li, K. S. M., Huang, Y., Tsoi, H.-W., Wong, B. H. L., Wong, S. S. Y., Leung, S.-Y., Chan, K.-H., & Yuen, K.-Y. (2005). Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(39). https://doi.org/10.1073/pnas.0506735102

Li, B., Deng, A., Li, K., Hu, Y., Li, Z., Xiong, Q., Liu, Z., Guo, Q., Zou, L., Zhang, H., Zhang, M., Ouyang, F., Su, J., Su, W., Xu, J., Lin, H., Sun, J., Peng, J., Jiang, H., Zhou, P., Hu, T., Luo, M., Zhang, Y., Zheng, H., Xiao, J., Liu, T., Che, R., Zeng, H., Zheng, Z., Huang, Y., Yu, J., Yi, L., Wu, J., Chen, J., Zhong, H., Deng, X.,

Kang, M., Pybus, O. G., Hall, M., Lythgoe, K. A., Li, Y., Yuan, J., He, J., & Lu, J. (2021). Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. *medRxiv*. https://doi.org/10.1101/2021.07.07.21260122

Li, Z., Tomlinson, A. C. A., Wong, A. H. M., Zhou, D., Desforges, M., Talbot, P. J., Benlekbir, S., Rubinstein, J. L., & Rini, J. M. (2019). The human coronavirus HCoV-229E S-protein structure and receptor binding. *eLife*, *8*, 512-530. https://doi.org/10.7554/eLife.51230

Liu, Y., Liu, J., Plante, K. S., Plante, J. A., Xie, X., Zhang, X., Ku, Z., An, Z., Scharton, D., Schindewolf, C., Menachery, V. D., Shi, P.-Y., & Weaver, S. C. (2021). The N501Y spike substitution enhances SARS-CoV-2 transmission. *bioRxiv : the preprint server for biology*. https://doi.org/10.1101/2021.03.08.434499

López-Fernández, H., Vieira, C. P., Ferreira, P., Gouveia, P., Fdez-Riverola, F., Reboiro-Jato, M., & Vieira, J. (2021). On the Identification of Clinically Relevant Bacterial Amino Acid Changes at the Whole Genome Level Using Auto-PSS-Genome. *Interdisciplinary Sciences: Computational Life Sciences*, *13*(2), 334-343. https://doi.org/10.1007/s12539-021-00439-2

López, G., Rojas, A., Tress, M., & Valencia, A. (2007). Assessment of predictions submitted for the CASP7 function prediction category. *Proteins*, *69 Suppl 8*, 165-174. https://doi.org/10.1002/prot.21651

Lv, Y., Li, Y., Yi, Y., Zhang, L., Shi, Q., & Yang, J. (2018). A Genomic Survey of Angiotensin-Converting Enzymes Provides Novel Insights into Their Molecular Evolution in Vertebrates. *Molecules (Basel, Switzerland)*, *23*(11). https://doi.org/10.3390/molecules23112923

MacLean, O. A., Lytras, S., Weaver, S., Singer, J. B., Boni, M. F., Lemey, P., Kosakovsky Pond, S. L., & Robertson, D. L. (2021). Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLOS Biology*, *19*(3). https://doi.org/10.1371/journal.pbio.3001115

Mamede, R. P., Pereira, M., & Simões, A. (2020). *Portugal : Rapid assessment of the impact of COVID-19 on the economy and labour market*. https://www.ilo.org/wcmsp5/groups/public/---ed_emp/documents/publication/wcms_749191.pdf

McIntosh, K., Dees, J. H., Becker, W. B., Kapikian, A. Z., & Chanock, R. M. (1967). Recovery in tracheal organ cultures of novel viruses from patients with respiratory disease. *Proceedings of the National Academy of Sciences of the United States of America*, *57*(4), 933-940. https://doi.org/10.1073/pnas.57.4.933

Meredith, R. W., Janečka, J. E., Gatesy, J., Ryder, O. A., Fisher, C. A., Teeling, E. C., Goodbla, A., Eizirik, E., Simão, T. L. L., Stadler, T., Rabosky, D. L., Honeycutt, R. L., Flynn, J. J., Ingram, C. M., Steiner, C., Williams, T. L., Robinson, T. J.,

Burk-Herrick, A., Westerman, M., Ayoub, N. A., Springer, M. S., & Murphy, W. J. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science*, *334*(6055). https://doi.org/10.1126/science.1211028

Mesel-Lemoine, M., Millet, J., Vidalain, P.-O., Law, H., Vabret, A., Lorin, V., Escriou, N., Albert, M. L., Nal, B., & Tangy, F. (2012). A Human Coronavirus Responsible for the Common Cold Massively Kills Dendritic Cells but Not Monocytes. *Journal of Virology*, *86*(14), 7577-7587. https://doi.org/10.1128/jvi.00269-12

Meyer, A., & Zardoya, R. (2003). Recent Advances in the (Molecular) Phylogeny of Vertebrates. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 311-338. https://doi.org/10.1146/annurev.ecolsys.34.011802.132351

Millet, J. K., & Whittaker, G. R. (2014). Host cell entry of Middle East respiratory syndrome coronavirus after two-step, furin-mediated activation of the spike protein. *Proceedings of the National Academy of Sciences*, *111*(42). https://doi.org/10.1073/pnas.1407087111

Millet, J. K., & Whittaker, G. R. (2015). Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis. *Virus Res*, *202*, 120-134. https://doi.org/10.1016/j.virusres.2014.11.021

Morfopoulou, S., Brown, J. R., Davies, E. G., Anderson, G., Virasami, A., Qasim, W., Chong, W. K., Hubank, M., Plagnol, V., Desforges, M., Jacques, T. S., Talbot, P. J., & Breuer, J. (2016). Human Coronavirus OC43 Associated with Fatal Encephalitis. *New England Journal of Medicine*, *375*(5), 497-498. https://doi.org/10.1056/NEJMc1509458

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function, and Bioinformatics*, *84*(S1), 4-14. https://doi.org/10.1002/prot.25064

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins*, *86 Suppl 1*(Suppl 1), 7-15. https://doi.org/10.1002/prot.25415

Müller, M. A., Meyer, B., Corman, V. M., Al-Masri, M., Turkestani, A., Ritz, D., Sieberg, A., Aldabbagh, S., Bosch, B. J., Lattwein, E., Alhakeem, R. F., Assiri, A. M., Albarrak, A. M., Al-Shangiti, A. M., Al-Tawfiq, J. A., Wikramaratna, P., Alrabeeah, A. A., Drosten, C., & Memish, Z. A. (2015). Presence of Middle East respiratory syndrome coronavirus antibodies in Saudi Arabia: a nationwide, cross-sectional, serological study. *Lancet Infect Dis*, *15*(5), 559-564. https://doi.org/10.1016/s1473-3099(15)70090-3

Murray, R. S., Brown, B., Brian, D., & Cabirac, G. F. (1992). Detection of coronavirus RNA and antigen in multiple sclerosis brain. *Annals of neurology*, *31*(5), 525-533. https://doi.org/10.1002/ana.410310511

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & Scheffler, K. (2013). FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Molecular Biology and Evolution*, *30*(5), 1196-1205. https://doi.org/10.1093/molbev/mst030

Myers, P., Espinosa, R., Parr, C. S., Jones, T., Hammond, G. S., & Dewey, T. A. (2021). *The Animal Diversity Web (online)*. https://animaldiversity.org

Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity*, *86*(6), 641-647. https://doi.org/10.1046/j.1365-2540.2001.00895.x

O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., Ruis, C., Abu-Dahab, K., Taylor, B., Yeats, C., du Plessis, L., Aanensen, D., Holmes, E., Pybus, O., & Rambaut, A. (2021). *pangolin: lineage assignment in an emerging pandemic as an epidemiological tool*. Retrieved 06/05/2021 from github.com/cov-lineages/pangolin

Pearson, C. A. B., Russell, T. W., Davies, N. G., Kucharski, A. J., Edmunds, W. J., Eggo, R. M., & CMMID COVID-19 working group. (2021). Estimates of severity and transmissibility of novel South Africa SARS-CoV-2 variant 501Y.V2.

Public Health England. (2020). *Investigation of novel SARS-CoV-2 variant: variant of concern 202012/01*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/950823/Variant_of_Concern_VOC_202012_01_Technical_Briefing_3_-_England.pdf

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

Riordan, J. F. (2003). Angiotensin-I-converting enzyme and its relatives. *Genome Biol*, *4*(8), 225. https://doi.org/10.1186/gb-2003-4-8-225

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*, *61*(3), 539-542. https://doi.org/10.1093/sysbio/sys029

Sabir, J. S., Lam, T. T., Ahmed, M. M., Li, L., Shen, Y., Abo-Aba, S. E., Qureshi, M. I., Abu-Zeid, M., Zhang, Y., Khiyami, M. A., Alharbi, N. S., Hajrah, N. H., Sabir, M. J., Mutwakil, M. H., Kabli, S. A., Alsulaimany, F. A., Obaid, A. Y., Zhou, B., Smith, D. K., Holmes, E. C., Zhu, H., & Guan, Y. (2016). Co-circulation of three camel

coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science*, *351*(6268), 81-84. https://doi.org/10.1126/science.aac8608

Salata, C., Calistri, A., Parolin, C., & Palù, G. (2020). Coronaviruses: a paradigm of new emerging zoonotic diseases. *Pathogens and Disease*, *77*(9). https://doi.org/10.1093/femspd/ftaa006

Sallard, E., Halloy, J., Casane, D., Decroly, E., & van Helden, J. (2021). Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review. *Environmental Chemistry Letters*. https://doi.org/10.1007/s10311-020-01151-1

Samara, E. M., & Abdoun, K. A. (2014). Concerns about Misinterpretation of Recent Scientific Data Implicating Dromedary Camels in Epidemiology of Middle East Respiratory Syndrome (MERS). *mBio*, *5*(4). https://doi.org/10.1128/mBio.01430-14

Schroth-Diez, B., Ludwig, K., Baljinnyam, B., Kozerski, C., Huang, Q., & Herrmann, A. (2000). The role of the transmembrane and of the intraviral domain of glycoproteins in membrane fusion of enveloped viruses. *Biosci Rep*, *20*(6), 571-595. https://doi.org/10.1023/a:1010415122234

Shang, J., Zheng, Y., Yang, Y., Liu, C., Geng, Q., Luo, C., Zhang, W., & Li, F. (2018). Cryo-EM structure of infectious bronchitis coronavirus spike protein reveals structural and functional evolution of coronavirus spike proteins. *PLOS Pathogens*, *14*(4). https://doi.org/10.1371/journal.ppat.1007009

Shang, J., Zheng, Y., Yang, Y., Liu, C., Geng, Q., Tai, W., Du, L., Zhou, Y., Zhang, W., & Li, F. (2018). Cryo-Electron Microscopy Structure of Porcine Deltacoronavirus Spike Protein in the Prefusion State. *Journal of Virology*, *92*(4). https://doi.org/10.1128/JVI.01556-17

Shulla, A., Heald-Sargent, T., Subramanya, G., Zhao, J., Perlman, S., & Gallagher, T. (2011). A Transmembrane Serine Protease Is Linked to the Severe Acute Respiratory Syndrome Coronavirus Receptor and Activates Virus Entry. *Journal of Virology*, *85*(2). https://doi.org/10.1128/JVI.02062-10

Sirotkin, K., & Sirotkin, D. (2020). Might SARS-CoV-2 Have Arisen via Serial Passage through an Animal Host or Cell Culture? *BioEssays*, *42*(10). https://doi.org/10.1002/bies.202000091

Smith, G. R., & Sternberg, M. J. (2002). Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol*, *12*(1), 28-35. https://doi.org/10.1016/s0959-440x(02)00285-3

Song, W., Gui, M., Wang, X., & Xiang, Y. (2018). Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLOS Pathogens*, *14*(8). https://doi.org/10.1371/journal.ppat.1007236

Steinhauer, D. A. (1999). Role of Hemagglutinin Cleavage for the Pathogenicity of Influenza Virus. *Virology*, *258*(1), 1-20. https://doi.org/https://doi.org/10.1006/viro.1999.9716

Su, S., Wong, G., Shi, W., Liu, J., Lai, A. C. K., Zhou, J., Liu, W., Bi, Y., & Gao, G. F. (2016). Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends in Microbiology*, *24*(6), 490-502. https://doi.org/10.1016/j.tim.2016.03.003

Suarez, D. L., Pantin-Jackwood, M. J., Swayne, D. E., Lee, S. A., DeBlois, S. M., & Spackman, E. (2020). Lack of Susceptibility to SARS-CoV-2 and MERS-CoV in Poultry. *Emerging infectious diseases*, *26*(12), 3074-3076. https://doi.org/10.3201/eid2612.202989

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A. J., Engelbrecht, S., Van Zyl, G., Preiser, W., Petruccione, F., Sigal, A., Hardie, D., Marais, G., Hsiao, M., Korsman, S., Davies, M.-A., Tyers, L., Mudau, I., York, D., Maslo, C., Goedhals, D., Abrahams, S., Laguda-Akingba, O., Alisoltani-Dehkordi, A., Godzik, A., Wibmer, C. K., Sewell, B. T., Lourenço, J., Alcantara, L. C. J., Pond, S. L. K., Weaver, S., Martin, D., Lessells, R. J., Bhiman, J. N., Williamson, C., & de Oliveira, T. (2020). Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv*. https://doi.org/10.1101/2020.12.21.20248640

Tortorici, M. A., & Veesler, D. (2019). Structural insights into coronavirus entry. *Advances in virus research*, *105*, 93-116. https://doi.org/10.1016/bs.aivir.2019.08.002

Tortorici, M. A., Walls, A. C., Lang, Y., Wang, C., Li, Z., Koerhuis, D., Boons, G.-J., Bosch, B.-J., Rey, F. A., de Groot, R. J., & Veesler, D. (2019). Structural basis for human coronavirus attachment to sialic acid receptors. *Nature Structural & Molecular Biology*, *26*(6), 481-489. https://doi.org/10.1038/s41594-019-0233-y

Towler, P., Staker, B., Prasad, S. G., Menon, S., Tang, J., Parsons, T., Ryan, D., Fisher, M., Williams, D., Dales, N. A., Patane, M. A., & Pantoliano, M. W. (2004). ACE2 X-ray structures reveal a large hinge-bending motion important for inhibitor binding and catalysis. *J Biol Chem*, *279*(17), 17996-18007. https://doi.org/10.1074/jbc.M311191200

Tyrrell, D. A., & Bynoe, M. L. (1965). Cultivation of a novel type of common-cold virus in organ cultures. *British medical journal*, *1*(5448), 1467-1470. https://doi.org/10.1136/bmj.1.5448.1467

UniProt: the universal protein knowledge base. (2017). *Nucleic Acids Res*, *45*(D1), D158-d169. https://doi.org/10.1093/nar/gkw1099

van der Hoek, L., Pyrc, K., Jebbink, M. F., Vermeulen-Oost, W., Berkhout, R. J. M., Wolthers, K. C., Wertheim-van Dillen, P. M. E., Kaandorp, J., Spaargaren, J., & Berkhout, B. (2004). Identification of a new human coronavirus. *Nature Medicine*, *10*(4), 368-373. https://doi.org/10.1038/nm1024

van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., Melquiond, A. S. J., van Dijk, M., de Vries, S. J., & Bonvin, A. M. J. J. (2016). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, *428*(4), 720-725. https://doi.org/10.1016/j.jmb.2015.09.014

Vlasak, R., Luytjes, W., Spaan, W., & Palese, P. (1988). Human and bovine coronaviruses recognize sialic acid-containing receptors similar to those of influenza C viruses. *Proceedings of the National Academy of Sciences*, *85*(12). https://doi.org/10.1073/pnas.85.12.4526

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

Walls, A. C., Tortorici, M. A., Bosch, B.-J., Frenz, B., Rottier, P. J. M., DiMaio, F., Rey, F. A., & Veesler, D. (2016). Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature*, *531*(7592), 114-117. https://doi.org/10.1038/nature16988

Walls, A. C., Tortorici, M. A., Frenz, B., Snijder, J., Li, W., Rey, F. A., DiMaio, F., Bosch, B.-J., & Veesler, D. (2016). Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. *Nature Structural & Molecular Biology*, *23*(10), 899-905. https://doi.org/10.1038/nsmb.3293

Walls, A. C., Tortorici, M. A., Snijder, J., Xiong, X., Bosch, B.-J., Rey, F. A., & Veesler, D. (2017). Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proceedings of the National Academy of Sciences*, *114*(42), 111-157. https://doi.org/10.1073/pnas.1708727114

Walls, A. C., Xiong, X., Park, Y.-J., Tortorici, M. A., Snijder, J., Quispe, J., Cameroni, E., Gopal, R., Dai, M., Lanzavecchia, A., Zambon, M., Rey, F. A., Corti, D., & Veesler, D. (2019). Unexpected Receptor Functional Mimicry Elucidates Activation of Coronavirus Fusion. *Cell*, *176*(5), 1026-1039. https://doi.org/10.1016/j.cell.2018.12.028

Warner, F. J., Smith, A. I., Hooper, N. M., & Turner, A. J. (2004). Angiotensin-converting enzyme-2: a molecular and cellular perspective. *Cell Mol Life Sci*, *61*(21), 2704-2713. https://doi.org/10.1007/s00018-004-4240-7

Wells, H. L., Letko, M., Lasso, G., Ssebide, B., Nziza, J., Byarugaba, D. K., Navarrete-Macias, I., Liang, E., Cranfield, M., Han, B. A., Tingley, M. W., Diuk-Wasser, M., Goldstein, T., Johnson, C. K., Mazet, J., Chandran, K., Munster, V. J., Gilardi, K.,

& Anthony, S. J. (2020). The evolutionary history of ACE2 usage within the coronavirus subgenus Sarbecovirus. *bioRxiv : the preprint server for biology*. https://doi.org/10.1101/2020.07.07.190546

Wilson, D. J., & McVean, G. (2006). Estimating Diversifying Selection and Functional Constraint in the Presence of Recombination. *Genetics*, *172*(3), 1411-1425. https://doi.org/10.1534/genetics.105.044917

Woo, P. C., Lau, S. K., Lam, C. S., Lau, C. C., Tsang, A. K., Lau, J. H., Bai, R., Teng, J. L., Tsang, C. C., Wang, M., Zheng, B. J., Chan, K. H., & Yuen, K. Y. (2012). Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. *J Virol*, *86*(7), 3995-4008. https://doi.org/10.1128/jvi.06540-11

Woo, P. C. Y., Lau, S. K. P., Chu, C.-m., Chan, K.-h., Tsoi, H.-w., Huang, Y., Wong, B. H. L., Poon, R. W. S., Cai, J. J., Luk, W.-k., Poon, L. L. M., Wong, S. S. Y., Guan, Y., Peiris, J. S. M., & Yuen, K.-y. (2005). Characterization and Complete Genome Sequence of a Novel Coronavirus, Coronavirus HKU1, from Patients with Pneumonia. *Journal of Virology*, *79*(2). https://doi.org/10.1128/JVI.79.2.884-895.2005

World Bank. (2020). Global Economic Prospects, June 2020. *Washington, DC: World Bank*. https://doi.org/10.1596/978-1-4648-1553-9

World Health Organization. (2015). *Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003*. Retrieved December 30, 2020 from https://www.who.int/publications/m/item/summary-of-probable-sars-cases-with-onset-of-illness-from-1-november-2002-to-31-july-2003

World Health Organization. (2019). *MERS Situation Update September 2019*. Retrieved January 02, 2021 from https://applications.emro.who.int/docs/EMROPub-MERS-SEP-2019-EN.pdf?ua=1&ua=1

Wu, K., Li, W., Peng, G., & Li, F. (2009). Crystal structure of NL63 respiratory coronavirus receptor-binding domain complexed with its human receptor. *Proceedings of the National Academy of Sciences*, *106*(47). https://doi.org/10.1073/pnas.0908837106

Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J.-J., Li, N., Guo, Y., Li, X., Shen, X., Zhang, Z., Shu, F., Huang, W., Li, Y., Zhang, Z., Chen, R.-A., Wu, Y.-J., Peng, S.-M., Huang, M., Xie, W.-J., Cai, Q.-H., Hou, F.-H., Chen, W., Xiao, L., & Shen, Y. (2020). Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature*, *583*(7815), 286-289. https://doi.org/10.1038/s41586-020-2313-x

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

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, *12*(1), 7-8. https://doi.org/10.1038/nmeth.3213

Yang, L., Wu, Z., Ren, X., Yang, F., Zhang, J., He, G., Dong, J., Sun, L., Zhu, Y., Zhang, S., & Jin, Q. (2014). MERS-related betacoronavirus in Vespertilio superans bats, China. *Emerging infectious diseases*, *20*(7), 1260-1262. https://doi.org/10.3201/eid2007.140318

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, *13*(5), 555-556. https://doi.org/10.1093/bioinformatics/13.5.555

Ye, Z.-W., Yuan, S., Yuen, K.-S., Fung, S.-Y., Chan, C.-P., & Jin, D.-Y. (2020). Zoonotic origins of human coronaviruses [Review]. *International Journal of Biological Sciences*, *16*(10), 1686-1697. https://doi.org/10.7150/ijbs.45472

Yim, H. E., & Yoo, K. H. (2008). Renin-Angiotensin System - Considerations for Hypertension and Kidney. *Electrolyte Blood Press*, *6*(1), 42-50.

Zhang, H., Wada, J., Hida, K., Tsuchiyama, Y., Hiragushi, K., Shikata, K., Wang, H., Lin, S., Kanwar, Y. S., & Makino, H. (2001). Collectrin, a Collecting Duct-specific Transmembrane Glycoprotein, Is a Novel Homolog of ACE2 and Is Developmentally Regulated in Embryonic Kidneys*. *Journal of Biological Chemistry*, *276*(20), 17132-17139. https://doi.org/10.1074/jbc.M006723200

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, *9*(1), 40. https://doi.org/10.1186/1471-2105-9-40

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, *33*(7), 2302-2309. https://doi.org/10.1093/nar/gki524

Zheng, W., Zhang, C., Bell, E. W., & Zhang, Y. (2019). I-TASSER gateway: A protein structure and function prediction server powered by XSEDE. *Future generations computer systems : FGCS*, *99*, 73-85. https://doi.org/10.1016/j.future.2019.04.011

Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., Wang, P., Liu, D., Yang, J., Holmes, E. C., Hughes, A. C., Bi, Y., & Shi, W. (2020). A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Current Biology*, *30*(11), 2196-2203. https://doi.org/10.1016/j.cub.2020.05.023

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-

Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., & Shi, Z.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270-273. https://doi.org/10.1038/s41586-020-2012-7

Zuiderweg, E. R. (2002). Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry*, *41*(1), 1-7. https://doi.org/10.1021/bi011870b

8. Supplementary data

Supplementary Table 1 – List of candidate species susceptible for SARS-CoV-2 infection obtained using the large scale predictive model. Damas score represents the results obtained by Damas et al. (2020). Species without this score were absent in their sample.

Class	Order	Species	Damas Score
Mammalia	Artiodactyla	Camelus bactrianus	Low
Mammalia	Artiodactyla	Camelus dromedarius	Low
Mammalia	Artiodactyla	Vicugna pacos	Medium
Mammalia	Artiodactyla	Bos indicus x Bos taurus	Medium
Mammalia	Artiodactyla	Bos mutus	Medium
Mammalia	Artiodactyla	Bos taurus	Medium
Mammalia	Artiodactyla	Bubalus bubalis	Medium
Mammalia	Artiodactyla	Capra hircus	Medium
Mammalia	Artiodactyla	Ovis aries	Medium
Mammalia	Artiodactyla	Sus scrofa	Low
Mammalia	Carnivora	Acinonyx jubatus	Medium
Mammalia	Carnivora	Felis catus	Medium
Mammalia	Carnivora	Lynx canadensis	Medium
Mammalia	Carnivora	Panthera pardus	Medium
Mammalia	Carnivora	Puma concolor	Medium
Mammalia	Carnivora	Panthera leo	-
Mammalia	Carnivora	Canis lupus dingo	Low
Mammalia	Carnivora	Vulpes vulpes	Low
Mammalia	Carnivora	Enhydra lutris kenyoni	Very low
Mammalia	Carnivora	Lontra canadensis	-
Mammalia	Carnivora	Mustela erminea	Very low
Mammalia	Carnivora	Mustela putorius furo	Very low
Mammalia	Carnivora	Neovison vison	Very low
Mammalia	Carnivora	Callorhinus ursinus	Very low
Mammalia	Carnivora	Eumetopias jubatus	Very low
Mammalia	Carnivora	Zalophus californianus	Very low
Mammalia	Carnivora	Mirounga leonina	-
Mammalia	Carnivora	Neomonachus schauinslandi	Very low
Mammalia	Carnivora	Odobenus rosmarus divergens	Very low
Mammalia	Carnivora	Phoca vitulina	Very low
Mammalia	Carnivora	Ursus arctos horribilis	Low
Mammalia	Cetacea	Balaenoptera a. scammoni	High
Mammalia	Cetacea	Delphinapterus leucas	High
Mammalia	Cetacea	Globicephala melas	High
Mammalia	Cetacea	Lagenorhynchus obliquidens	High
Mammalia	Cetacea	Lipotes vexillifer	High
Mammalia	Cetacea	Monodon monoceros	High
Mammalia	Cetacea	Neophocaena a. asiaeorientalis	High
Mammalia	Cetacea	Orcinus orca	High
Mammalia	Cetacea	Phocoena sinus	-
Mammalia	Cetacea	Physeter catodon	Medium

Continues next page

Class	Order	Species	Damas Score
Mammalia	Cetacea	Tursiops truncatus	High
Mammalia	Chiroptera	Eptesicus fuscus	Very low
Mammalia	Chiroptera	Phyllostomus discolor	Very low
Mammalia	Chiroptera	Pteropus alecto	Low
Mammalia	Chiroptera	Pteropus vampyrus	Low
Mammalia	Chiroptera	Rousettus aegyptiacus	Low
Mammalia	Diprotodontia	Phascolarctos cinereus	Very low
Mammalia	Diprotodontia	Vombatus ursinus	Very low
Mammalia	Lagomorpha	Oryctolagus cuniculus	Medium
Mammalia	Lagomorpha	Ochotona princeps	Low
Mammalia	Perissodactyla	Equus asinus	Low
Mammalia	Perissodactyla	Equus caballus	Low
Mammalia	Perissodactyla	Equus przewalskii	Low
Mammalia	Perissodactyla	Ceratotherium simum simum	Low
Mammalia	Pholidota	Manis javanica	Very low
Mammalia	Primates	Aotus nancymaae	Medium
Mammalia	Primates	Cercocebus atys	Very high
Mammalia	Primates	Chlorocebus sabaeus	Very high
Mammalia	Primates	Gorilla gorilla gorilla	Very high
Mammalia	Primates	Homo sapiens	Very high
Mammalia	Primates	Macaca fascicularis	Very high
Mammalia	Primates	Macaca mulatta	Very high
Mammalia	Primates	Macaca nemestrina	Very high
Mammalia	Primates	Mandrillus leucophaeus	Very high
Mammalia	Primates	Microcebus murinus	Low
Mammalia	Primates	Nomascus leucogenys	Very high
Mammalia	Primates	Pan paniscus	Very high
Mammalia	Primates	Pan troglodytes	Very high
Mammalia	Primates	Papio anubis	Very high
Mammalia	Primates	Piliocolobus tephrosceles	Very high
Mammalia	Primates	Rhinopithecus roxellana	Very high
Mammalia	Primates	Saimiri boliviensis boliviensis	Medium
Mammalia	Primates	Sapajus apella	Medium
Mammalia	Primates	Theropithecus gelada	Very high
Mammalia	Primates	Trachypithecus francoisi	-
Mammalia	Primates	Hylobates moloch	-
Mammalia	Primates	Callithrix jacchus	Medium
Mammalia	Primates	Pongo abelii	Very high
Mammalia	Primates	Otolemur garnettii	Very low
Mammalia	Primates	Propithecus coquereli	High
Mammalia	Proboscidea	Loxodonta africana	Low
Mammalia	Rodentia	Peromyscus leucopus	Low
Mammalia	Rodentia	Peromyscus maniculatus bairdii	Medium
Mammalia	Rodentia	Cricetulus griseus	High
Mammalia	Rodentia	Mesocricetus auratus	Medium
Mammalia	Rodentia	Microtus ochrogaster	Low
Mammalia	Rodentia	Mastomys coucha	Very low

Continues next page

Class	Order	Species	Damas Score
Mammalia	Rodentia	Mus caroli	Very low
Mammalia	Rodentia	Mus pahari	Very low
Mammalia	Rodentia	Rattus rattus	-
Mammalia	Rodentia	Mus musculus	Very low
Mammalia	Rodentia	Grammomys surdaster	Very low
Mammalia	Rodentia	Nannospalax galili	High
Mammalia	Rodentia	Marmota flaviventris	Medium
Mammalia	Rodentia	Jaculus jaculus	Very low
Mammalia	Rodentia	Heterocephalus glaber	Medium
Mammalia	Rodentia	Ictidomys tridecemlineatus	Medium
Mammalia	Rodentia	Urocitellus parryii	Medium
Mammalia	Rodentia	Dipodomys ordii	Medium
Mammalia	Scandentia	Tupaia chinensis	Very low
Mammalia	Sirenia	Trichechus manatus latirostris	Low
Aves	Passeriformes	Camarhynchus parvulus	Very low
Aves	Passeriformes	Cyanistes caeruleus	Very low
Aves	Passeriformes	Ficedula albicollis	Very low
Aves	Passeriformes	Serinus canaria	Very low
Aves	Passeriformes	Sturnus vulgaris	Very low
Aves	Passeriformes	Zonotrichia albicollis	Very low
Reptilia	Squamata	Zootoca vivipara	Very low

Supplementary Table 2 – Alignment by MEGAX of the protein pattern region of the list of CoVs used by Wells et al. (2020). Red and blue entries represent species able and unable to use ACE2, respectively, as an entry receptor. Pink and light blue entries represent species with no evidence of ACE2 use that clustered with red and blue entries, respectively, when only using RBD sequences (Wells et al., 2020). The pattern discovered is composed of the sites in green, with the position number corresponding to the SARS-CoV-2 reference S protein.

CoVs species	SARS-CoV-2 reference S protein position													
	488	489	490	491	492	493	494	495	496	497	498	499	500	501
YP 009724390.1 S glycoprotein SARS-CoV-2	C	Y	F	P	L	Q	S	Y	G	F	Q	P	T	N
QHR63300.2 Bat CoV RaTG13	.	.	Y	.	.	Y	R	.	.	.	Y	.	.	D
ATO98205.1 Bat SARS-like CoV	.	.	W	.	.	N	D	.	.	.	Y	I	.	.
YP 009825051.1 SARS CoV Tor2	.	.	W	.	.	N	D	.	.	.	Y	T	.	T
ATO98218.1 Bat SARS-like CoV	.	.	W	.	.	N	D	.	.	.	F	T	.	.
AGZ48806.1 Bat SARS-like CoV RsSHC014	.	.	N	.	.	R	P	.	.	.	F	T	.	A
ATO98132.1 Bat SARS-like CoV	.	.	N	.	.	R	P	.	.	.	F	T	.	A
ATO98157.1 Bat SARS-like CoV	.	.	N	.	.	R	P	.	.	.	F	T	.	A
ALK02457.1 SARS-like CoV WIV16	.	.	W	.	.	N	D	.	.	.	Y	I	.	.
AGZ48831.1 Bat SARS-like CoV WIV1	.	.	W	.	.	N	D	.	.	.	Y	I	.	.
AHX37558.1 Rhinolophus affinis CoV	.	.	W	.	.	N	D	.	.	.	Y	T	.	.
AY304486 SARS-SZ3	.	.	W	.	.	K	D	.	.	.	Y	T	.	S
EPI ISL 410721 P2S	.	.	.	.	.	.	.	.	.	.	H	.	.	.
AGZ48818.1 Bat SARS-like CoV Rs3367	.	.	W	.	.	N	D	.	.	.	Y	I	.	.
ATO98231.1 Bat SARS-like CoV	.	.	W	.	.	N	D	.	.	.	F	T	.	.
QDF43825.1 CoV BtRs-BetaCoV/YN2018B	.	.	W	.	.	N	D	.	.	.	F	T	.	.

Continues next page

CoVs species	SARS-CoV-2 reference S protein position													
	488	489	490	491	492	493	494	495	496	497	498	499	500	501
QIQ54048.1 Pangolin CoV	.	.	Y	.	.	E	R	.	.	.	H	.	.	T
ABD75323.1 Bat SARS CoV Rf1/2004	G	V	R	T	.	S	T	.	D	.	N	Q	N	V
AIA62290.1 BtRf-BetaCoV/HeB2013	G	V	R	T	.	S	T	.	D	.	N	Q	Y	V
ABG47060.1 Bat CoV 273/2005	N	G	R	T	.	S	T	.	D	.	N	Q	N	V
AIA62277.1 BtRf-BetaCoV/JL2012	G	V	L	T	.	S	T	.	D	.	N	Q	N	V
AVP78031.1 Bat SARS-like CoV	G	V	R	T	.	S	T	.	D	.	N	.	N	V
AVP78042.1 Bat SARS-like CoV	G	V	R	T	.	S	T	.	D	.	N	.	N	V
AIA62330.1 BtRs-BetaCoV/YN2013	G	V	R	T	.	S	T	.	D	.	N	.	N	V
ATO98145.1 Bat SARS-like CoV	G	V	R	T	.	S	T	.	D	.	N	.	N	V
AGC74176.1 Bat CoV Cp/Yunnan2011	G	V	R	T	.	S	T	.	D	.	Y	.	S	V
ATO98120.1 Bat SARS-like CoV	G	V	R	T	.	S	T	.	D	.	Y	.	N	V
ATO98169.1 Bat SARS-like CoV	G	V	R	T	.	S	T	.	D	.	Y	.	.	V
ATO98108.1 Bat SARS-like CoV	G	V	R	T	.	S	T	.	D	.	Y	.	.	V
AAZ67052.1 Bat SARS CoV Rp3/2004	G	V	R	T	.	S	T	.	D	.	Y	.	S	V
ABG47069.1 Bat CoV 279/2005	G	V	Y	T	.	S	T	.	D	.	Y	.	S	I
AGC74165.1 Bat CoV Rp/Shaanxi2011	G	V	Y	T	.	S	T	.	D	.	Y	.	S	V
ADE34823.1 Bat SARS CoV HKU3-13	G	V	Y	T	.	S	T	.	D	.	N	.	N	V
ADE34766.1 Bat SARS CoV HKU3-8	G	V	Y	T	.	S	T	.	D	.	N	.	N	V
AID16716.1 Bat SARS-like CoV	G	V	Y	T	.	S	T	.	D	.	N	.	N	V
AIA62320.1 BtRs-BetaCoV/GX2013	G	V	Y	T	.	S	T	.	D	.	N	.	N	V
ATO98181.1 Bat SARS-like CoV	G	V	Y	T	.	S	T	.	D	.	N	.	N	V
AIA62310.1 BtRs-BetaCoV/HuB2013	G	V	Y	T	.	S	T	.	D	.	N	.	N	V
ADK66841.1 Bat CoV BM48-31/BGR/2008	.	.	K	.	.	A	.	.	.	.	T	Q	S	S
QTJ30153.1 Bat CoV	.	.	E	.	.	K	.	.	.	.	T	.	.	V
QTJ30135.1 Bat CoV	.	.	E	.	.	K	.	.	.	.	T	.	.	V
QTJ30144.1 Bat CoV	.	.	E	.	.	K	.	.	.	.	T	.	.	V
APO40579.1 S protein SARS-related CoV	.	.	E	.	.	K	.	.	.	.	T	.	.	V
AIA62300.1 BtRf-BetaCoV/SX2013	G	V	R	T	.	S	T	.	D	.	N	Q	Y	V
ARI44809.1 Bat CoV	G	V	R	T	.	S	T	.	D	.	N	Q	Y	V
ANA96027.1 Bat CoV	G	V	L	T	.	S	T	.	D	.	N	Q	N	V
ASO66810.1 Bat CoV	G	V	L	T	.	S	T	.	D	.	N	Q	N	V
AKZ19076.1 Bat SARS-like CoV YNLF 31C	G	A	R	T	.	S	T	.	D	.	N	Q	N	V
KZ19087.1 Bat SARS-like CoV YNLF 34C	G	A	R	T	.	S	T	.	D	.	N	Q	N	V
ARI44799.1 Bat CoV	G	V	R	T	.	S	T	.	D	.	N	.	Y	V
ARI44804.1 Bat CoV	G	V	R	T	.	S	T	.	D	.	N	.	Y	V
ARO76382.1 SARS-related CoV	G	V	R	T	.	S	T	.	D	.	N	.	N	V
ANA96090.1 partial Bat CoV	G	V	R	T	.	S	T	.	D	.	Y	.	.	V
ACU31032.1 SARS CoV Rs 672/2006	G	V	R	T	.	S	T	.	D	.	Y	.	N	V
QDF43835.1 CoV BtRs-BetaCoV/YN2018D	G	V	R	T	.	S	T	.	D	.	Y	.	N	V
ATO98193.1 Bat SARS-like CoV	G	V	R	T	.	S	T	.	D	.	Y	.	.	V
QDF43830.1 CoV BtRs-BetaCoV/YN2018C	G	V	R	T	.	S	T	.	D	.	Y	.	.	V
ABD75332.1 Bat SARS CoV Rm1/2004	G	V	Y	T	.	S	T	.	D	.	Y	.	S	I
ADE34812.1 Bat SARS CoV HKU3-12	G	V	Y	T	.	S	T	.	D	.	N	.	N	V
QDF43820.1 CoV BtRs-BetaCoV/YN2018A	G	V	Y	T	.	S	T	.	D	.	N	.	N	V
QDF43815.1 CoV BtRl-BetaCoV/SC2018	G	V	Y	T	.	S	T	.	D	.	N	.	N	V

Supplementary Table 3 – The TM-Scores of pairing the RBM region of the S protein of SARS-CoV-2 and species of CoVs separated by their ability to bind to ACE2. Scores normalized for the size of SARS-CoV-2's S protein.

Species unable to use ACE2	TM-score	Species able to use ACE2	TM-score
AIA62310.1 HuB2013	0,549	QHR63300.2 RaTG13	0,799
AID16716.1 Longquan 140	0,599	QDF43825.1 YN2018B	0,918
ADE34755.1 HKU3-7	0,655	BAC81390.1 SARS TWS	0,959
ADE34823.1 HKU3-13	0,525	ATO98205.1 Rs4874	0,914
QDF43820.1 YN2018A	0,596	AGZ48806.1 RsSHC014	0,931
AIA62320.1 GX2013	0,583	AGZ48831.1 WIV1	0,937
AAZ67052.1 Rp3/2004	0,621	ALK02457.1 WIV16	0,753
ABD75323.1 Rf1/2004	0,509	AY304486 SARS-SZ3	0,943
AIA62277.1 JL2012	0,594	EPI ISL 410721 P2S	0,881
ADK66841.1 BM48-31	0,858	ATO98218.1 Rs7327	0,939
AGC74165.1 Shaanxi2011	0,666	ATO98132.1 Rs4084	0,941
		ATO98157.1 Rs4231	0,937

Supplementary Table 4 – List of PSS obtained using codeML and FUBAR ordered by the frequency the appear in the 100 random samples. The frequency is normalized for 100 samples.

codeML		FUBAR	
Reference S protein position	**Frequency per 100 random samples**	**Reference S protein position**	**Frequency per 100 random samples**
677	0,32	677	0,82
681	0,31	501	0,80
5	0,29	484	0,77
501	0,24	5	0,75
484	0,23	681	0,60
152	0,21	477	0,58
452	0,18	26	0,34
477	0,18	18	0,29
716	0,15	138	0,28
732	0,15	494	0,27
26	0,14	732	0,26
253	0,14	95	0,25
13	0,13	253	0,24
95	0,13	452	0,20
570	0,13	701	0,19
614	0,13	716	0,19
982	0,13	80	0,16
1118	0,13	1191	0,16
1106	0,12	675	0,14
1129	0,12	859	0,14
1191	0,12	67	0,09
478	0,11	520	0,09
701	0,11	98	0,08

Continues next page

codeML		FUBAR	
Reference S protein position	**Frequency per 100 random samples**	**Reference S protein position**	**Frequency per 100 random samples**
18	0,10	222	0,08
138	0,09	655	0,08
494	0,08	152	0,07
20	0,07	570	0,07
80	0,07	1118	0,07
1176	0,07	1176	0,07
67	0,06	20	0,06
190	0,06	157	0,05
417	0,06	1228	0,05
675	0,06	52	0,04
812	0,06	215	0,04
1078	0,06	688	0,04
52	0,05	1078	0,04
655	0,05	1111	0,04
780	0,05	13	0,03
859	0,05	22	0,03
1027	0,05	27	0,03
54	0,04	49	0,03
98	0,04	54	0,03
222	0,04	143	0,03
258	0,04	614	0,03
382	0,04	679	0,03
520	0,04	769	0,03
888	0,04	982	0,03
1228	0,04	12	0,02
22	0,03	29	0,02
157	0,03	142	0,02
215	0,03	190	0,02
522	0,03	255	0,02
950	0,03	417	0,02
12	0,02	478	0,02
27	0,02	572	0,02
142	0,02	640	0,02
173	0,02	845	0,02
176	0,02	936	0,02
177	0,02	957	0,02
228	0,02	1020	0,02
384	0,02	1070	0,02
583	0,02	1129	0,02
640	0,02	1264	0,02
688	0,02	9	0,01
769	0,02	33	0,01
836	0,02	69	0,01
845	0,02	72	0,01
854	0,02	75	0,01

Continues next page

codeML		FUBAR	
Reference S protein position	**Frequency per 100 random samples**	**Reference S protein position**	**Frequency per 100 random samples**
879	0,02	111	0,01
939	0,02	141	0,01
1040	0,02	173	0,01
1111	0,02	176	0,01
1122	0,02	178	0,01
1164	0,02	214	0,01
1202	0,02	219	0,01
1263	0,02	221	0,01
6	0,01	226	0,01
14	0,01	228	0,01
21	0,01	250	0,01
32	0,01	251	0,01
41	0,01	252	0,01
46	0,01	254	0,01
49	0,01	257	0,01
72	0,01	308	0,01
111	0,01	323	0,01
145	0,01	367	0,01
150	0,01	382	0,01
153	0,01	498	0,01
154	0,01	519	0,01
163	0,01	522	0,01
175	0,01	544	0,01
178	0,01	583	0,01
180	0,01	622	0,01
181	0,01	638	0,01
189	0,01	641	0,01
198	0,01	653	0,01
206	0,01	698	0,01
214	0,01	704	0,01
219	0,01	778	0,01
220	0,01	780	0,01
221	0,01	834	0,01
226	0,01	943	0,01
227	0,01	950	0,01
249	0,01	1027	0,01
250	0,01	1101	0,01
251	0,01	1117	0,01
254	0,01	1144	0,01
255	0,01	1167	0,01
257	0,01	1203	0,01
259	0,01	1219	0,01
263	0,01	1263	0,01
298	0,01		
307	0,01		

Continues next page

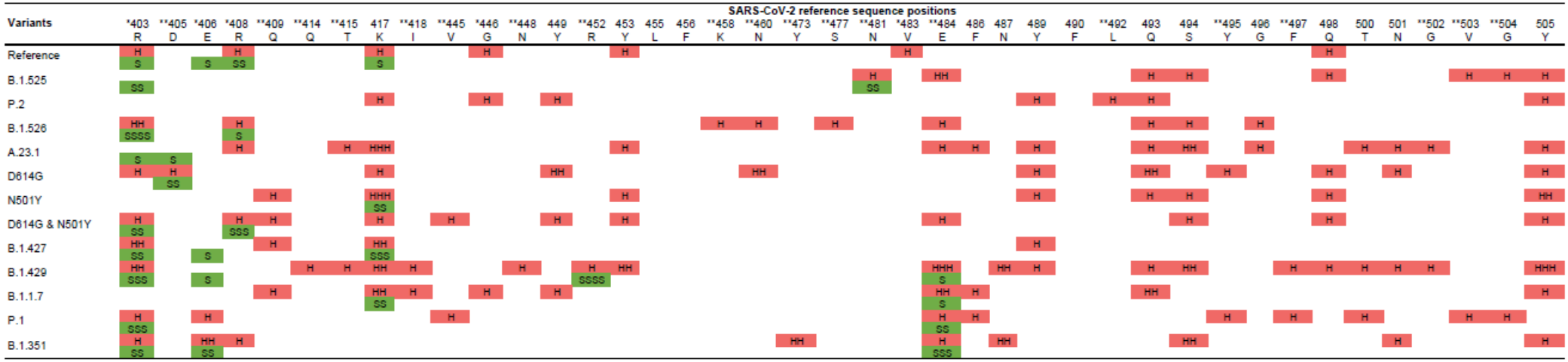
codeML		FUBAR	
Reference S protein position	**Frequency per 100 random samples**	**Reference S protein position**	**Frequency per 100 random samples**
309	0,01		
315	0,01		
327	0,01		
338	0,01		
344	0,01		
362	0,01		
403	0,01		
427	0,01		
439	0,01		
440	0,01		
448	0,01		
476	0,01		
486	0,01		
519	0,01		
529	0,01		
554	0,01		
556	0,01		
561	0,01		
584	0,01		
608	0,01		
613	0,01		
635	0,01		
638	0,01		
654	0,01		
676	0,01		
678	0,01		
679	0,01		
683	0,01		
698	0,01		
704	0,01		
733	0,01		
761	0,01		
772	0,01		
778	0,01		
809	0,01		
813	0,01		
826	0,01		
835	0,01		
838	0,01		
871	0,01		
892	0,01		
929	0,01		
957	0,01		
958	0,01		
981	0,01		

Continues next page

codeML		FUBAR	
Reference S protein position	**Frequency per 100 random samples**	**Reference S protein position**	**Frequency per 100 random samples**
1006	0,01		
1014	0,01		
1018	0,01		
1050	0,01		
1063	0,01		
1072	0,01		
1087	0,01		
1099	0,01		
1101	0,01		
1103	0,01		
1104	0,01		
1120	0,01		
1153	0,01		
1159	0,01		
1162	0,01		
1168	0,01		
1219	0,01		
1229	0,01		
1230	0,01		
1231	0,01		
1235	0,01		
1237	0,01		
1248	0,01		
1255	0,01		

Supplementary Table 5 – Representation of the residues involved in the hydrogen bonds and salt bridges of the S protein (A) and hACE2 (B), given by PDBePISA, for SARS-CoV-2 VOIs and VOCs. Pink colour represents hydrogen bonds, green colour salt bridges. The number of letters in a given site equals to the number of bonds a given residues has. The sites presented correspond: interfacing sites in 100% of the variants, sites that present a HB or SB with the reference sequence (*) and sites having a HB or SB in a variant but not in the reference sequence (**).

A



B

