Virus Research 302 (2021) 198472

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Positive selection as a key player for SARS-CoV-2 pathogenicity: Insights into ORF1ab, S and E genes

Mohamed Emam [a], Mariam Oweda [a], Agostinho Antunes [b,c,*], Mohamed El-Hadidi [a]

[a] Bioinformatics group, Center for Informatics Sciences (CIS), Nile University, Giza, Egypt
[b] CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos, s/n, 4450-208 Porto, Portugal
[c] Department of Biology, Faculty of Sciences, University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal

ARTICLE INFO

Keywords:
Positive selection
SARS-COV-2
Maximum likelihood
Pathogenicity
Drug targeting

ABSTRACT

The human β-coronavirus SARS-CoV-2 epidemic started in late December 2019 in Wuhan, China. It causes Covid-19 disease which has become pandemic. Each of the five-known human β-coronaviruses has four major structural proteins (E, M, N and S) and 16 non-structural proteins encoded by ORF1a and ORF1b together (ORF1ab) that are involved in virus pathogenicity and infectivity. Here, we performed detailed positive selection analyses for those six genes among the four previously known human β-coronaviruses and within 38 SARS-CoV-2 genomes to assess signatures of adaptive evolution using maximum likelihood approaches. Our results suggest that three genes (E, S and ORF1ab genes) are under strong signatures of positive selection among human β-coronavirus, influencing codons that are located in functional important protein domains. The E protein-coding gene showed signatures of positive selection in two sites, Asp 66 and Ser 68, located inside a putative transmembrane α-helical domain C-terminal part, which is preferentially composed by hydrophilic residues. Such Asp and Ser sites substitutions (hydrophilic residues) increase the stability of the transmembrane domain in SARS-CoV-2. Moreover, substitutions in the spike (S) protein S1 N-terminal domain have been found, all of them were located on the S protein surface, suggesting their importance in viral transmissibility and survival. Furthermore, evidence of strong positive selection was detected in three of the SARS-CoV-2 nonstructural proteins (NSP1, NSP3, NSP16), which are encoded by ORF1ab and play vital roles in suppressing host translation machinery, viral replication and transcription and inhibiting the host immune response. These results are insightful to assess the role of positive selection in the SARS-CoV-2 encoded proteins, which will allow to better understand the virulent pathogenicity of the virus and potentially identifying targets for drug or vaccine strategy design

Abbreviation

SARS-CoV-2	The severe acute respiratory syndrome coronavirus-2
MERS-CoV	Middle East respiratory syndrome coronavirus
WHO	World Health Organization
ICTV	International Committee on Taxonomy of Viruses
E	Small envelope protein
M	Matrix protein
N	Nucleocapsid protein
S	Spike protein
ACE2	Angiotensin converting enzyme 2
HBC	Human β-coronavirus
CDS	Coding sequences
UTR	Untranslated region
MSA	Multiple sequence alignment
ECDF	Empirical cumulative function
AICc	Akaike information criterion correction
LRT	likelihood ratio tests
BEB	Bayes Empirical Bayes
NEB	Naïve Empirical Bayes
NSPs	Non-structural proteins
HVR	Hypervariable region

[*] Corresponding author at: CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos, s/n, 4450-208 Porto, Portugal.
E-mail address: aantunes@ciimar.up.pt (A. Antunes).

https://doi.org/10.1016/j.virusres.2021.198472
Received 22 April 2021; Received in revised form 2 June 2021; Accepted 3 June 2021
Available online 10 June 2021

1. Introduction

The severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) epidemic emerged in early December 2019 in Wuhan, Hubei Province, China (Wang et al., 2020a,b). The disease that is caused by this virus has been termed Covid-19 (the '19' in Covid-19 stands for the year 2019) by the World Health Organization (WHO) on February 19, 2020. The '19' in COVID-19 stands for the year 2019. Taxonomically, SARS-CoV-2 belongs to the existing species *Severe acute respiratory syndrome-related coronavirus* as determined by the *Coronaviridae* Study Group of the International Committee on Taxonomy of Viruses (ICTV) on February 2, 2020. The species is a member of the genus *Betacoronavirus* and the family *Coronaviridae* (Gorbalenya et al., 2020). Since last December, COVID-19 has rapidly spread across different areas in China and subsequently many countries causing pandemic. The major clinical symptoms of the disease in patients are fever, pneumonia, dry cough, headache, and dyspnea. The progression of the disease may result in progressive respiratory failure due to alveolar damage and even may lead death (Li et al., 2020). The virus is highly transmissible among humans and infected individuals may shed the virus efficiently in the first week of infection when they are asymptomatic or show mild symptoms (Wölfel et al., 2020). SARS-CoV-2 is also possibly transmissible to pangolins (Choo et al., 2020), ferrets and cats (Shi et al. 2020); with cats being highly susceptible to the virus air born infection. As of 18 of April 2021, the global total confirmed COVID-19 cases is 140, 322,903 and deaths is 3,003,794 https://coronavirus.jhu.edu/map.html.

Other members of the genus *Betacoronavirus* that infect humans include SARS-CoV-1, Middle East respiratory syndrome coronavirus (MERS-CoV) and two other viruses, HCoV -OC43 and HCoV-HKU1. SARS-CoV-1 emerged in 2002 and MERS-COV emerged in 2012 with limited transmission from human to human (Tang et al., 2015, Song et al., 2019). Both viruses caused severe illness with fatality rate of approximately 9 and 36%, respectively. HCoV-OC43 and HCoV-HKU1 are considered the second most common cause of the common cold and their infection may cause respiratory tract illness (Al-Khannaq et al., 2016; Cui et al., 2018; Li et al., 2020). The genomes of these viruses are single-stranded positive-sense RNAs whose size varies from 26,000 to 32,000 nucleotides (nt) with six to eleven open reading frames (ORFs) (Song et al., 2019), which encode accessory proteins, major structural proteins and non-structural proteins (NSPS) (Cui et al., 2018).

The RNA genome of SARS-CoV-2 has 29, 811 nt that contain 14 ORFs encoding 27 proteins (Wu et al., 2020). The 3'-terminus of the genome contains eight accessory proteins and four structural proteins. The structural proteins are: small envelope protein (E), matrix protein (M), nucleocapsid protein (N), which binds to the viral RNA genome and the spike protein (S) located at the surface of the virus envelope. The S protein binds to a receptor termed angiotensin converting enzyme 2 (ACE2) to enter into host cells and determine host tropism (Li, 2016; Zhu et al., 2018). There are 16 NSPs located at the 5'-terminus of the genome. The pp1ab and pp1a proteins are encoded by the orf1ab and orf1a genes, respectively. Together, they comprise 15 NSPs including from NSP1 to NSP10 and NSP12 to NSP16.

Comparative analysis of genomic data demonstrated that SARS-CoV-2 evolved naturally and it is not man-made construct biological agent (Anderson et al., 2020). In a phylogenetic network analysis of SARS-CoV-2 were found two central variants observed and termed as A and B lineages. A.1 lineage was the Primary outbreak in Washington State, USA and B.1 with B.2 lineage were comprised the large Italian outbreak (Rambaut et al., 2020).

Previous studies have shown the extent of molecular divergence between SARS-CoV-2 and other related coronaviruses. It was found that the nucleotide divergence at synonymous sites between SARS-CoV-2 and other coronaviruses such as SARSr-CoV and RaTG13 was much higher than previously expected (Tang et al., 2020). Selective constraints during the evolution of SARS-CoV-2 and related coronaviruses indicate strong negative selection on the nonsynonymous sites. Therefore, although these coronaviruses coding sequences were generally under very strong negative selection, positive selection was also responsible for the evolutionary shaping of the protein sequences (Angeletti, et al., 2020; Tang et al., 2020). The genes that are involved in functional innovation often show the footprints of positive selection through high ratios of nonsynonymous to synonymous nucleotide substitutions (Yang, 2007; Nielsen, 2005; Philip et al., 2012). Hence, it is essential to perform an in-depth comprehensive positive selection analysis on the functional sites. In this study, we focused on positive selection analysis of SARS -CoV-2 structural genes among Human β-coronavirus (HBC) species and within 36 genomes of SARS-CoV-2, on both coding and non-coding regions. This work provides insights into the key role of positive selection on the recent pathogenicity of the virus and its transmission pattern among humans as well as into E, S and ORF1ab protein, which can identify potential drug targets or vaccine strategy.

2. Materials and methods

2.1. Sequencing data retrieval

All coding sequences (CDS) and the non-coding regions (3'-UTR and 5′-UTR) were downloaded from the NCBI virus portal (https://www.ncbi.nlm.nih.gov/genome/virus). Information about genes and accession numbers of the 36 SARS-CoV-2 genomes used in this study can be found in supplementary Table S1. The reference sequences of the coding regions of five HBC species were retrieved from the NCBI RefSeq database (OLeary et al., 2015), each species represented by three strains. For each viral genome, the information of the noncoding regions (3'-UTR and 5′-UTR) was extracted from 36 SARS-CoV2 genomes, 50 SARS CoV genomes, 35 HCoV-HKUI genomes, 50 HCoV-OC43 genomes and 50 MERS CoV genomes. Accession numbers of these genomes are listed in supplementary Table S1, Supplementary Material.

2.2. Substitution rate of the coding sequences

Estimation of the positively selected sites was implemented through multiple sequence alignments (MSA) by using SEAVIEW v4 (Gouy et al., 2009). The coding sequences were translated to amino acids, aligned using MUSCLE (Edgar, 2004) and further back-translated to nucleotides, then the MSAs were filtered with GBLOCKS (Castresana, 2000) using the relaxed parameters (Talavera and Castresana, 2007) to avoid misaligned positions and eliminate false-positive hits. JMODELTEST v2.1.10 (Darriba et al., 2012) was used for maximum likelihood ratio test to select the best-fit model and then Akaike information criterion correction (AICc) was used for model ranking. Construction of phylogenetic gene-based trees were built using PhyML v3.0 (Guindon et al., 2009) under the best-fit model (Tables S2 and S3). The data set contained six refined MSAs between HBC CDS (E gene, M gene, N gene ORF1a, ORF1ab and S gene) with an average length 40,296 bps and 10 refined MSAs within SARS CoV2 strains (E gene, M gene, N gene, ORF1ab, S gene, ORF3a, ORF6, ORF7, ORF8 and ORF10) with an average length 2915 bps. The ratio between nonsynonymous (dN) and synonymous (dS) substitution, known as omega (ω) were estimated using the maximum-likelihood method CODEML in PAML v4.6 (Yang, 2007). Genes were compared to a neutrally evolving model, where ω is equal to one. This value can be considered as evidence of positive selection when the value of $\omega > 1$, or as purifying selection when the value of $\omega < 1$. Estimation of dN /dS ratio for each amino acid site was obtained using three different models (7, 8 and 8a). Equilibrium codon frequencies of the model were used as free parameters (CodonFreq = 2). The Model 7 (M7, beta) is a null model contains the sites-classes which are lower or equal to the neutrality and Model 8 (M8, beta + $\omega > 1$) as an alternative model was used to observe differences over sites through a beta distribution, whereas M8 only contains the sites-classes that above neutrality. As model 8 allows positive selection along the alignment, we compared model 8 pairwise against a stricter model which is M7, using likelihood ratio tests (LRT). Each calculation of the LRT corresponds to 2 × [*lnL* (alternative model)−*lnL* (*null* model)] (or LRT = 2 × (Δ*lnL*)). We performed a comparison between models M8 and M8a to identify deviations from neutrality, focusing on testing whether sites belonging to a site-class with a $d_N/d_S >$ 1 are evolving differently from near neutrality ($d_N/d_S \approx 1$). The LRTs obtained from each pairwise comparison between model M7 versus M8 and M8 versus M8a were used to extract the *P*-value from the chi-square distribution with two degrees of freedom in the case of M7 versus M8 and one degree of freedom in the case of M8 versus M8a, the *P*-value was adjusted using FDR correction method (Benjamini and Hochberg, 1995), genes were considered to be under positive selection in case of having a significant difference in both model comparisons with adjusted *p-value* lower than 0.05.

2.3. Substitution rate of the non-coding sequences

Multiple sequence alignment (MSA) were built using SEAVIEW v4 (Gouy et al 2009). Both 3'-UTR and 5'-UTR alignments were built using MUSCLE (Edgar, 2004). JMODELTEST v2.1.10 (Darriba et al., 2012) was used for maximum likelihood ratio test to select the best-fit model and then we used Akaike information criterion correction (AICc) for model ranking. Construction of phylogenetic gene-based trees were built using PhyML v3.0 (Guindon et al., 2009) under the best-fit model. The data set contained ten refined MSAs of the five HBC 3'-UTR and 5'-UTR (five 3'-UTR and five 5'-UTR). PhyloP wig-scores analysis was performed using PHAST (Hubisz et al., 2010) to measure the evolutionary conservation and acceleration at individual alignment sites (positive scores for conservation sites and the negative scores for acceleration sites). The Mann–Whitney *U* test *P* values and the empirical cumulative function (ECDF) of 5'-UTR and 3'-UTR PhyloP wig-scores were performed using R studio vR1.1.2.5. By obtaining multiple random samples of 3'-UTR and 5'-UTR wig-scores value for each analyzed nucleated position, we performed a validated comparison between the five HBC, the results of the comparisons between five viruses 3'-UTR and 5'-UTR were tested using the Mann–Whitney U values.

3. Result

The genomic evidence reveals a signature of strong positive selection sites for E, S and ORF1ab genes among HBC species. When both MSAs and gene-based trees were used as input for CODEML analysis, M7 versus M8 comparison was significantly more adjusted in five genes, although while using M8 versus M8a (the strict model comparison), we observed four genes which showed that the site class was significantly above neutrality. E gene, S gene, ORF1a and ORF1ab genes LRT tests comparisons have significant differences, M7 versus M8 chi-square showed statistically significant adjusted FDR correction for multiple comparisons *P-values* of $P < 0.01$ (E gene), $P <$ 3.364e-07 (S gene), $P < 1.182e\text{-}11$ (ORF1a) and $P < 2.595e\text{-}17$ (ORF1ab). The chi-square adjusted *P*-value for M8 versus M8a showed values of $P < 5.633e\text{-}15$ (E gene), $P < 0.004$ (S gene), $P < 0.00$ (ORF1a) and $P < 0.039$ (ORF1ab) (Table 1).

According to the Bayes Empirical Bayes (BEB) analysis only three genes have posterior probability above 80% and posterior probability above 90 % in the Naïve Empirical Bayes (NEB) analysis, which are E, S and ORF1ab. For the E gene, we found two codons under positive selection with their posterior probability equal or over 95% for each codon, residues position and their posterior amino acids probability (Table 1). Regarding the S gene, we found three codons under positive selection and four codons in the ORF1ab under positive selection (residues position and their amino acids substitutions (Table 1). By mapping E protein against the domain database using the NCBI domain blast (Marchler-Bauer et al., 2014), we found both residues (66 Asparagine and 68 Serine) are in the SARS-CoV-2_E domain with E-value *2.02e-24* (Figure 1). The SARS-CoV-2_E domain is involved in the virus morphogenesis and assembly (Raamsman et al., 2000); it acts as a viroporin and induce self-assembly in the host membranes, which plays a central role in ion transport with poor selectivity through forming homopentameric protein-lipid pores. The domains of the spike protein

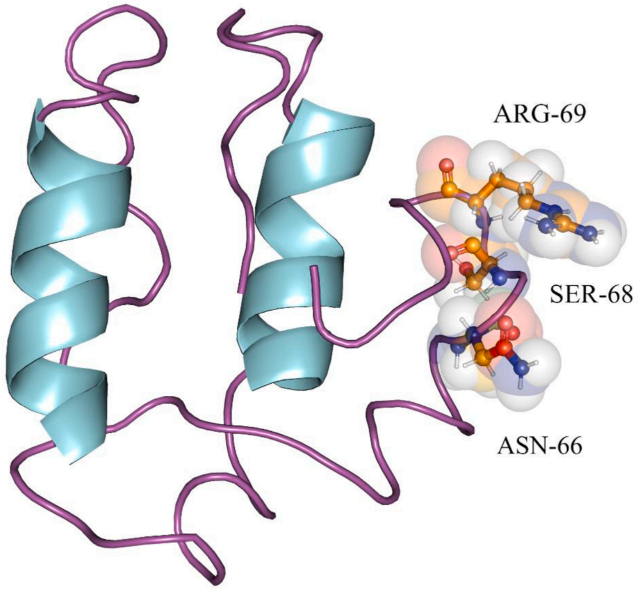

Fig. 1. I-Tasser model of the SARS-COV-2 E protein (QHD43418). positively selected residues with a $P < 0.05$ are shown as transparent spheres and are marked by the corresponding labels.

Table 1
E protein, S protein and ORF1ab protein positively selected sites, residue positions and their amino acids change and their posterior probability for each codon between HBC their accession number are: (NC_045512.2: SARS COV2 Wuhan-Hu-1, NC_004718.3: SARS coronavirus Tor2, FJ882947.1: SARS coronavirus wtic-MB, FJ882926.1: SARS coronavirus ExoN1, NC_006577.2: Human coronavirus HKU1, KF430201.1: Human coronavirus HKU1-18, KF686342.1: Human coronavirus HKU1-11, NC_006213.1: Human coronavirus OC43, KF530099.1: Human coronavirus OC43-971-5, MK303621.1: Human coronavirus OC43 MDS4, NC_019843.3: Middle East respiratory syndrome coronavirus(MERS), MN723542.1: MERS Riyadh-KSA-036D1N, MG757605.1: MERS KSA-036D1N).

Protein	Position	Positively selected AA	Substitution	Posterior probability (BEB)	Posterior probability (NEB)
E Protein	66	ASN (N)	VLA (V), LYS (K) and SER (S)	0.948	0.997
	68	SER (S)	PRO (P) and GLU (E)	0.978	0.993
S Protein	26	PRO (P)	ARG (R), PHE (F), SER (S), LYS (K) and ASN (N)	0.875	0.990
	148	ASN (N)	LYS (K) and PRO (P)	0.851	0.984
	153	MET (M)	ARG (R), PHY (F), TYR (Y) and THR (T)	0.802	0.902
ORF1ab Protein					
Nsp1	138	ALA (A)	ILE (I), CYS (C), ARG (A) and TYR (Y)	0.842	0.970
Nsp3	196	MET(M)	LEU (L), VAL (V) and GLU (E)	0.823	0.939
	1229	VAL (V)	GLU (E), GLY (G), SER(S) and Thr (T)	0.807	0.923
Nsp16	216	ARG (R)	Lys (K) and SER (S)	0.820	0.939

were identified using the protein families database (Pfam), we found that all of the three positively selected sites (Pro 26, Asn 148 and Met 153) were located in the S1 N-terminal domain with E*-value 5E.-71* (Figure 2).

However, we did not find significant differences between M7 vs M8 and M8 vs M8a models (Table 2) regarding the coding sequences within the 36 SARS- CoV-2 strain present in this study, but the non-coding sequence of SARS- CoV- 2 showed a high evolutionary rate. The ECDF comparison (Figure 3) between the five HBC showed an acceleration in the 3'-UTR and 5'-UTR in SARS -CoV- 2 with significant differences (Mann–Whitney *U* test, $P < 0.01$) at the lower rank (higher acceleration, $P < 0.01$). As the non-coding part (3'-UTR and 5'-UTR) is accumulative for the mutations, we can consider the high acceleration of SARS- CoV- 2 as evidence of a higher evolutionary rate (Machado et al., 2016) (the pairwise Mann–Whitney *U* test for both 3'-UTR and 5′-UTR is presented in Tables S4 and S5).

4. Discussion

Previous studies confirmed that coronavirus proteins vary in size, and this can be described as pleomorphic. Interestingly, even in the conserved set of components between the homologous structural proteins, less than 30% in amino acid identity is observed. Hence, we performed a detailed positive-selection analysis for functional sites of six genes among five HBC and ten genes within 36 SARS -CoV- 2 strains to understand the effect of natural selection in the powerful infectivity of SARS- CoV-2. Our findings reveal signatures of strong positive selection of three genes: E gene, S gene and ORF1ab between HBC.

4.1. E gene

E gene translated into a small pentameric structure protein that delimits an ion conductive pore, which plays a crucial role in virus-host interaction (Torres et al., 2006). In the previous studies, recombinant CoVs lacking the E protein result in significantly decrease on the virus titres, reduced maturation, or yield propagation incompetent progeny (Dewald Schoeman and Fielding, 2019). The E protein of SARS-COV2 is highly similar to the SARS-CoV E protein, which has one putative transmembrane α-helical hydrophobic domain, 20–30 amino acids long, flanked by N-terminus (short amino acids sequence <10 amino acids) and a longer C-terminus tail, both more hydrophilic (Torres et al., 2006). According to NCBI domain blast, both sites 66 Asn and 68 Ser of the E protein are within an alpha-helical transmembrane domain C-terminal part. We found that site 66 substitutions from Ser, Val and Lys into Asn in SARS- CoV-2 and site 68 substitutions from Glu and Pro into Ser in SARS- CoV-2 (Supplementary Fig. 1 S1), which either increase or maintain the polarity of the C-terminal part of the domain. Such substitutions into highly hydrophilic amino acids inside the C-terminal may enhance the stability of the E protein, which increases SARS COV2 production, maturation and pathogenicity.

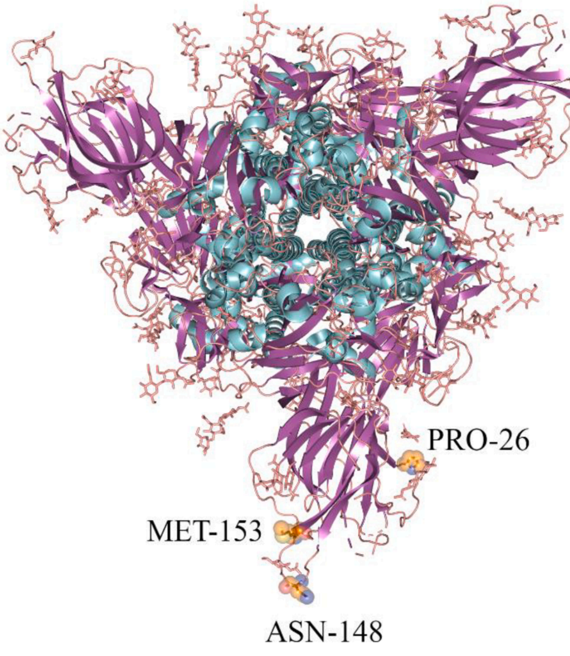

Fig. 2. PDB structure of S protein (6XR8). positively selected residues with a *P < 0.05* are shown as transparent spheres and are marked by the corresponding labels.

4.2. Spike gene

The SARS-COV-2 spike glycoprotein (S) is the largest structural protein of the virus (Pillay, 2020), it plays a vital role in the viral infection through its binding with the human ACE2 receptor to initiate the viral entry (Lan et al., 2020), spike protein binding affinity to ACE2 is correlated with the replication rate in different species and also with viral contagiousness and severity (Guan et al., 2003a,b; Li et al., 2005; Wan et al, 2020). The spike protein is composed of two main subunits; S1 which is responsible for ACE2 receptor binding via its receptor binding domain and S2 which mediates viral and cellular membranes fusion (Walls et al. 2020). In our study we found three positively selected sites in the extracellular N-terminal domain (NTD) of the S1 subunit, which are Pro 26, Asn 148 and Met 153. The pro 26 is located in a loop structure of S1 NTD (Fig. 2), this site lies within P25PA sequon which corresponds to N29YT sequon is SARS-COV, in SARS-COV this sequon; N29YT, was found to be glycosylated, however, in SARS-COV-2 it is no longer glycosylated (Walls et al., 2020), this could suggest a probable differentiating mechanism between SARS-COV-2 and SARS-COV. The asparagine 148 resides at the β turns of s1 subunit surface, Asn is more favorable on the protein surfaces due to its polarity (Kyte and Doolittle, 1982) in comparison with proline in both SARS and MERS (Supplementary Fig. 2 S2, Fig. 3 S3). The last site Met 153 lies on the β sheets of the S1 subunit, the methionine is preferable inside the β sheets structure (Bhattacharjee and Biswas, 2010). Moreover, it can act as a ligand for metal ions (Betts and Russell, 2007).

4.3. ORF1ab

The ORF1ab represents two-thirds of the viral genome that encodes the polyprotein 1ab (pp1ab) that is cleaved into 16 non-structural proteins (NSPs), which are involved in viral transcription and replication (Brian and Baric, 2005). Our analysis revealed that three of these (NSPs) contain strong positively selected sites: NSP1, NSP3 and NSP16.

NSP1 is one of the first proteins to be expressed after the viral infection to inhibit the host translation machinery through multiple steps of binding with 40S and 80S ribosomal complexes, blocking the mRNAs entry location and suppressing the host antiviral mechanisms, which rely on the expression of host immune factors such as interferons (Lokugamage et al., 2012; Thoms et al., 2020). Moreover, the NSP1-40S ribosomal complex initiate endonucleolytic activity to degrade the host mRNAs, however, the viral genes continue to be efficiently translated due to NSP1 and the viral genes 5′ untranslated region (UTR) interaction (Huang et al., 2011; Schubert et al., 2020). NSP1 is composed of N-terminal domain followed by a flexible unstructured linker, and the C-terminal domain which binds with the 40S mRNA entry site, due to the linker flexibility, the N-terminal domain could sample a space of ~ 60 Å from its point of attachment. However, the linker structure is still unresolved (Schubert et al., 2020; Thoms et al., 2020). The Ala 138 residue substitution is located in the flexible linker of the NSP1, Ala is more flexible than other COVs amino acids in the same position (refer to the alignment figure Supplementary s1) (Huang and Nau, 2003; Koča

Table 2
The CODEML output contains the LRT result for M7 vs M8 and M8 vs M8a models and the *P*-value for each of the studied genes. HBC (Human β-coronavirus) and SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2).

Gene	Model 7 (lnL) null model	Model 8 (lnL) alt model	Model 8a (lnL) null model	LRT (M7 vs M8)	*p*-value (adjusted)	LRT (M8 vs M8a)	*p*-value (adjusted)
M gene (HBC)	-3194.123	-3194.123	-3193.271	0	1	-1.704336	1
N gene (HBC)	-6751.231	-6744.739	-6743.957	12.983	0.0022	-1.564974	1
ORF1a (HBC)	-62127.61	-62101.35	-62109.86	52.519172	1.182e-11	17.016658	0.00011
ORF1ab (HBC)	-95269.129	-95229.14	-95231.61	79.963936	2.595e-17	4.92846	0.03962
S gene (HBC)	-19554.79	-19539.19	-19543.94	31.196074	3.364e-07	9.493334	0.004124
E gene (HBC)	-1198.212	-1193.354	-1225.632	9.715472	0.00932	64.554818	5.633e-15
E gene (SARS COV 2)	-299.1647	-298.6350	-298.8755	1.059568	1	0.481106	0.48792
M gene SARS COV 2)	-904.370	-903.7801	-903.780	1.179642	1	-0.000224	1
ORF1ab SARS COV 2)	-28191.49555	-28189.43693	-28190.8	4.11724	1	2.780354	0.954
S gene SARS COV 2)	-5042.16	-5042.16	-5042.16	-4.4E-05	1	-0.000	1
N gene SARS COV 2)	-1711.67	-1711.67	-1711.67	-0.000674	1	-0.0004	1
ORF3a SARS COV 2)	-1114.01	-1113.97	-1113.97	0.06619	1	-0.003	1
ORF6 SARS COV 2)	-224.847	-224.847	-224.8472	0.000682	1	-0.000	1
ORF7 SARS COV 2)	-478.682	-478.682	-478.683	-2E-06	1	0.0010	1
ORF8 SARS COV 2)	-488.540	-487.950	-488.495	1.179918	1	1.09062	1
ORF10 SARS COV 2)	-148.138	-148.138	-148.138	0	1	0	1

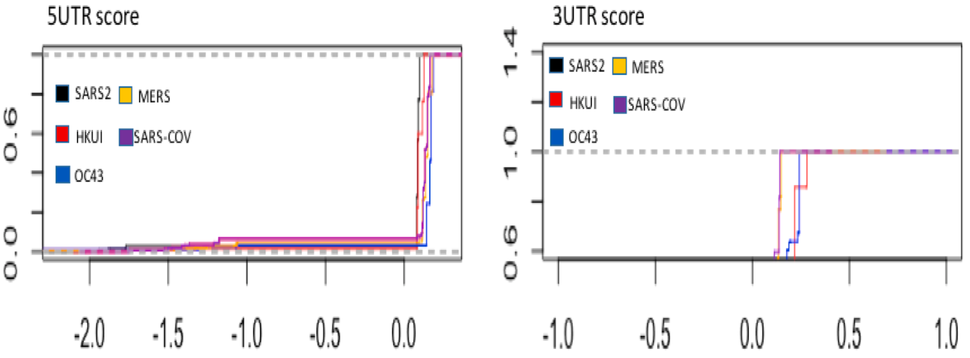

Fig. 3. ECDF for comparison among SARS-CoV2, SARS CoV, HCoV-HKUI, HCoV-OC43 and MERS CoV (3’-UTR and 5′-UTR).

et al., 1994), thus, we can interpret that this substitution may increase the flexibility of the linker.

Nsp3 is the largest non-structural protein in the genome of coronavirus, containing multiple functional domains that are required for coronavirus replication and blocking host innate immune response (Lei et al., 2018). Here we found two sites under positive selection within different two domains: Met 196 and Val 1229 (Figure 4) in the Glu-rich acidic region and beta coronavirus-specific marker (βSM) domain, respectively (Ong et al., 2020a,b).

Glu-rich acidic region comprises more than 35% Glu and 10% Asp residues, it is also known as the hypervariable region (HVR) due to its non-conserved amino-acid sequence (Neuman, 2016), till now the function of this region is still unknown. In general, Glu/Asp rich proteins mainly involved DNA/ RNA mimicry, protein−protein interactions and metal-ion binding (Chou and Wang, 2015). The Met 196 is an amphipathic amino acid that substituted into Lue and Val which are non-polar amino acids in HCoV-HKU1 and HCoV-OC43, respectively, and also substituted into Glu which is polar amino acids in SARS and MERS. Glu, Lue and Val are more abundant in the Glu-rich acidic region in comparison with Met (Chou and Wang, 2015). However, the ability of Met to donate a methyl group (National Center for Biotechnology Information 2020) could suggest a relevancy of this position.

The second substitution Val 1229 lies within betacoronavirus-specific marker domain (βSM), an intrinsically disordered region with low conservation (Lei et al., 2018; Ong et al., 2020a,b). The role of the βSM in viral pathogenesis is still unknown. The gene that codes the SARS-CoV domain βSM could not be expressed in E-coli, suggesting that βSM is a non-enzymatic domain (Neuman et al., 2008). The Val 1229 is located in in the βSM alpha helix structure of NSP3 I-TASSER model, in spite of the Val weakly destabilizing the alpha helix structure it was found to be more favored than Gly and Thr in HCoV-OC43 and SARS, respectively, but less favored than Glu in HCoV-HKU1 (Supplementary Fig. 4 S4) (Nick Pace and Martin Scholtz, 1998).

NSP16 plays a critical role in viral transcription and replication; during RNA synthesis. NSP16 adds a cap structure to the newly synthesized viral mRNAs, ensuring their efficient translation (Bouvet et al., 2010). NSP16 negatively regulates innate immunity to promote viral proliferation through interferon inhibition (Shi et al., 2019). In all SARS CoV, MERS and HCoV-OC43, Arg 216 residue replaced Lys in the same position of NSP16 (Fig. 5, Supplementary Fig. 5 S5). Both amino acids have very similar characteristics. However, arginine can bind via multiple hydrogen bonds with the negatively charged groups on phosphates structure such as in RNA more than lysine does.

Recent studies that have analyzed SARS-CoV-2 mutations, discovered that among all mutations, C to T exchanges existed in preponderance of more than 50% and revealed that hypermutations of C > T are most likely resulting from the APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) deamination in RNA editing (Di

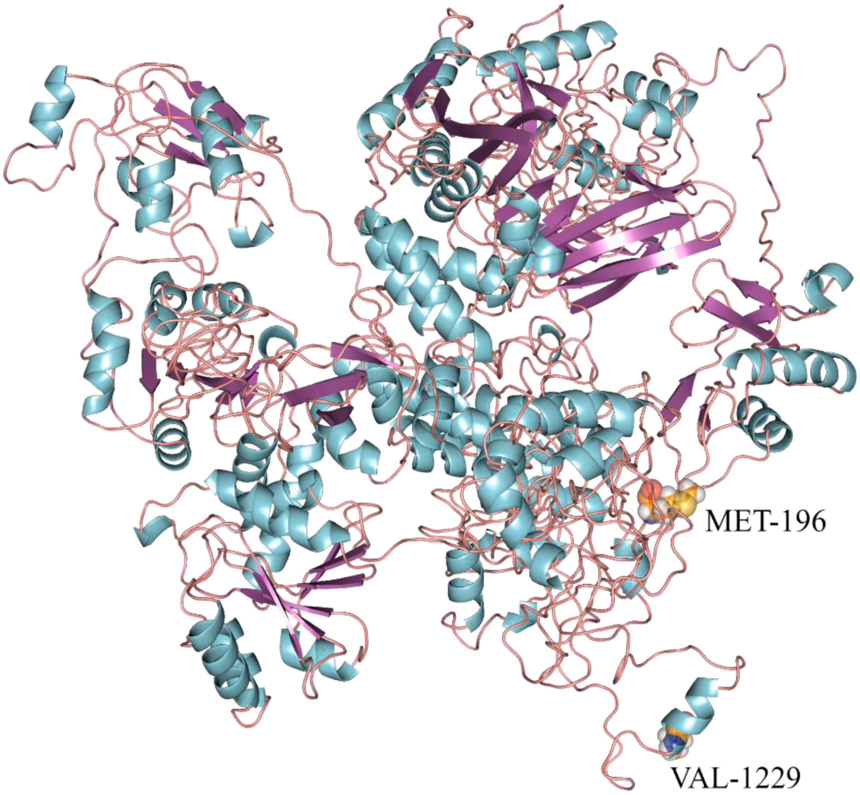

Fig. 4. I-Tasser model of the SARS-COV-2 NSP3 (QHD43415_3). positively selected residues with a $P < 0.05$ are shown as transparent spheres and are marked by the corresponding labels.

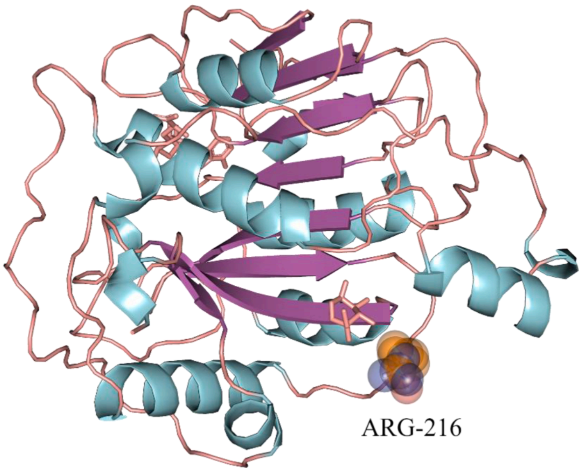

Fig. 5. PDB crystal structure of NSP16 (6w75*). positively selected residues with a $P < 0.05$ are shown as transparent spheres and are marked by the corresponding labels. * This accession number contain Crystal Structure of NSP16 - NSP10 Complex however in this figure we present the NSP16 as it is the main focus.

Giorgio et al., 2020). This finding is similar to the exchange preferences in our study results as we found five positivly selected sites having C to T mutations (namley position 68 SER on E gene, position 148 ASN on S gene, postion 138 ALA (A) on Nsp1, position 1229 VAL (V) on Nsp3, and poisiton 216 ARG (R) on Nsp16 protein). This large proportion of C > T mutations in a host APOBEC-like context, provides evidence for a potent host-driven antiviral editing mechanism against the pathogencity of SARS-CoV-2 to improve cellular defense functions (Wang et al., 2020a,b; Simmonds, 2020).

We did not find evidence of positive selection within SARS COV2 genomes with our method, this result support another recent study findings, which was evaluating SARS COV2 recombination, they did not find genes under positive selection within SARS COV2, but they found patterns of purifying selection pressure in some parts of the genome, including the E and M genes, as well as the partial ORF1a and ORF1b genes, which plays an important role in cross-species transmission (Li et al., 2020b).

In addition, to further evidence of positive selection between HBC in our results, we evaluated non-coding parts (3'-UTR and 5′-UTR) among five HBC through the PhyloP score, showing a higher acceleration rate in both (3'-UTR and 5′-UTR) of SARS- CoV-2 providing further evidence of a consistent higher evolutionary rate concordant with the presence of positive selection in coding regions (Tables S4 and S5).

5. Conclusion

Our results suggest that S, E and ORF1ab genes are under strong signatures of positive selection among human β-coronaviruses, affecting codons that reside in functionally important protein domains. Overall, most of the substitutions increase protein structure stability. The positively selected sites in these proteins could justify some clinical features of SARS-CoV-2 compared with other human β-coronaviruses. Sites undergoing an amino acid change are insightful to highlight relevant functionally important proteins of the SARS-CoV-2 that are essential for the mechanism of viral replication, transcription and evading the host's antiviral immunity. While the current literature contains a huge flow of data about SARS-CoV-2 mutagenesis and variants, limited insights were retrieved regarding the impact of those mutations on biological processes and viral pathogenicity. Here we shed light on the role of these proteins and their associated mutations on the viral pathogenicity and host biological processes. Furthermore, our findings could reveal valuable information useful for potential drug and vaccines development.

Author contributions

M.E-H. supervised the study. M.E-H. and A.A. equally participated in the design, genetic analyses, drafting, and coordination of the study. M. E. performed the phylogenetic and evolutionary analyses. M.O. participated in modeling and results interpretation. M.E. and M.O. drafted the manuscript. M.E-H. and A.A. revised the manuscript. All authors read the manuscript, and approved to be co-authors on the manuscript and have a substantial contribution to the manuscript.

Declaration of Competing Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Acknowledgements

A.A. was partially supported by the Strategic Funding UIDB/04423/2020 and UIDP/04423/2020 through national funds provided by the Fundação para a Ciência e Tecnologia (FCT) and the European Regional Development Fund (ERDF) in the framework of the program PT2020, by the European Structural and Investment Funds (ESIF) through the Competitiveness and Internationalization Operational Program - COMPETE 2020 and by National Funds through the FCT under the project PTDC/CTA-AMB/31774/2017 (POCI-01-0145-FEDER/031774/2017).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.virusres.2021.198472.

References

Al-Khannaq, M.N., Ng, K.T., Oong, X.Y., Pang, Y.K., Takebe, Y., Chook, J.B., et al., 2016. Molecular epidemiology and evolutionary histories of human coronavirus OC43 and HKU1 among patients with upper respiratory tract infections in Kuala Lumpur, Malaysia. Virol. J. 13, 33. https://doi.org/10.1186/s12985-016-0488-4.

Anderson, K.G., Rambaut, A., Lipkin, W.I, Holmes, E.C., Garry, R.F., 2020. The proximal origin of SARS-CoV-2. Nat. Med. 26, 450–452. https://doi.org/10.1038/s41591-020-0820-9.

Angeletti, S., Benvenuto, D., Bianchi, M., Giovanetti, M., Pascarella, S., Ciccozzi, M., 2020. COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. J. Med. Virol. 92, 584–588. https://doi.org/10.1002/jmv.25719.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Betts, M.J., Russell, R.B., 2007. Amino-acid properties and consequences of substitutions. Bioinform. Genet. 311–342. https://doi.org/10.1002/9780470059180.ch13.

Bhattacharjee, N., Biswas, P., 2010. Position-specific propensities of amino acids in the beta-strand. BMC Struct. Biol. 10, 29. https://doi.org/10.1186/1472-6807-10-29.

Bouvet, M., Debarnot, C., Imbert, I., Selisko, B., Snijder, E.J., Canard, B., et al., 2010. In vitro reconstitution of SARS-coronavirus mRNA cap methylation. PLoS Pathog. 6 https://doi.org/10.1371/journal.ppat.1000863.

Brian, D.A., Baric, R.S., 2005. Coronavirus genome structure and replication. Curr. Top. Microbiol. Immunol. Coronavirus Replic. Reverse Genet. 1–30. https://doi.org/10.1007/3-540-26765-4_1.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

Choo, S.W., Zhou, J., Tian, X., Zhang, S., Qiang, S., O'Brien, S.J., Tan, K.Y., Platto, S., Koepfli, K.P., Antunes, A., Sitam, F.T., 2020. Are pangolins scapegoats of the COVID-19 outbreak-CoV transmission and pathology evidence? Conserv. Lett. 13 (6), e12754. https://doi.org/10.1111/conl.12754.

Chou, C.-C., Wang, A.H.-J., 2015. Structural D/E-rich repeats play multiple roles especially in gene regulation through DNA/RNA mimicry. Mol. Biosyst. 11, 2144–2151. https://doi.org/10.1039/c5mb00206k.

Cui, J., Li, F., Shi, Z.-L., 2018. Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. 17, 181–192. https://doi.org/10.1038/s41579-018-0118-9.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D, 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9. https://doi.org/10.1038/nmeth.2109, 772–772.

Di Giorgio, S., Martignano, F., Torcia, M.G., Mattiuz, G., Conticello, S.G., et al., 2020. Evidence for RNA editing in the transcriptome of 2019 novel coronavirus. Sci. Adv. https://doi.org/10.1126/sciadv.abb5813, 2020, 6, eabb5813.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340.

Gorbalenya, A.E., Baker, S.C., Baric, R.S., de Groo, R.J., Drosten, C., Gulyaeva, A.A, Haagmans, B.L., Lauber, C., Lauber, A.M., Neuman, B.W., et al., 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. Nat. Microbiol. 5, 536–544. https://doi.org/10.1038/s41564-020-0695-z.

Gouy, M., Guindon, S., Gascuel, O., 2009. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221–224. https://doi.org/10.1093/molbev/msp259.

Guan, Y., Zheng, B., He, Y., Liu, X., Zhuang, Z., Cheung, C., Luo, S., Li, P., Zhang, L., Guan, Y., et al., 2003a. Isolation and characterization of viruses related to the SARS coronavirus from animals in Southern China. Science 302, 276–278. https://doi.org/10.1126/science.1087139.

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., et al., 2003b. Isolation and characterization of viruses related to the SARS coronavirus from animals in Southern China. Science 302, 276–278. https://doi.org/10.1126/science.1087139.

Guindon, S., Delsuc, F., Dufayard, J.-F., Gascuel, O., 2009. Estimating Maximum Likelihood Phylogenies with PhyML. Methods in Molecular Biology Bioinformatics for DNA Sequence Analysis 537, 113–137. https://doi.org/10.1007/978-1-59745-251-9_6.

Huang, C., Lokugamage, K., Rozovics, J., Narayanan, K., Semler, B., Makino, S., 2011. SARS coronavirus nsp1 protein induces template-dependent endonucleolytic cleavage of mRNAs: viral mRNAs are resistant to nsp1-induced RNA cleavage. PLoS Pathog. 7, e1002433 https://doi.org/10.1371/journal.ppat.1002433.

Huang, F., Nau, W., 2003. A conformational flexibility scale for amino acids in peptides. Angew. Chem. Int. Ed. 42, 2269–2272. https://doi.org/10.1002/anie.200250684.

Hubisz, M.J., Pollard, K.S., Siepel, A., 2010. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief. Bioinform. 12, 41–51. https://doi.org/10.1093/bib/bbq072.

Koča, J., Kříž, Z., Carlsen, P., 1994. Computer study of conformational flexibility of 20 common amino acids. J. Mol. Struct. THEOCHEM 306, 157–164. https://doi.org/10.1016/0166-1280(94)80036-7.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132. https://doi.org/10.1016/0022-2836(82)90515-0.

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

Lei, J., Kusov, Y., Hilgenfeld, R., 2018. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. Antiviral Res. 149, 58–74. https://doi.org/10.1016/j.antiviral.2017.11.001.

Li, F., 2016. Structure, function, and evolution of coronavirus spike proteins. Annu. Rev. Virol. 3, 237–261. https://doi.org/10.1146/annurev-virology-110615-042301.

Li, W., Zhang, C., Sui, J., Kuhn, J., Moore, M., Luo, S., Wong, S., Huang, I., Xu, K., Vasilieva, N., et al., 2005. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. EMBO J. 24, 1634–1643. https://doi.org/10.1038/sj.emboj.7600640.

Li, X., Song, Y., Wong, G., Cui, J., 2020a. Bat origin of a new human coronavirus: there and back again. Sci. China Life Sci. 63, 461–462. https://doi.org/10.1007/s11427-020-1645-7.

Lokugamage, K., Narayanan, K., Huang, C., Makino, S., 2012. Severe acute respiratory syndrome coronavirus protein nsp1 is a novel eukaryotic translation inhibitor that represses multiple steps of translation initiation. J. Virol. 86, 13598–13608. https://doi.org/10.1128/jvi.01958-12.

Machado, J.P., Philip, S., Maldonado, E., O'Brien, S.J., Johnson, W.E., Antunes, A, 2016. Positive selection linked with generation of novel mammalian dentition patterns. Genome Biol. Evol. 8, 2748–2759. https://doi.org/10.1093/gbe/evw200.

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., et al., 2014. CDD: NCBIs conserved domain database. Nucleic Acids Res. 43 https://doi.org/10.1093/nar/gku1221.

National Center for Biotechnology Information (2020). PubChem Compound Summary for CID 6137, Methionine. Retrieved August 7, 2020 from https://pubchem.ncbi.nlm.nih.gov/compound/Methionine.

Neuman, B.W., 2016. Bioinformatics and functional analyses of coronavirus nonstructural proteins involved in the formation of replicative organelles. Antiviral Res. 135, 97–107. https://doi.org/10.1016/j.antiviral.2016.10.005.

Neuman, B.W., Joseph, J.S., Saikatendu, K.S., Serrano, P., Chatterjee, A., Johnson, M.A., et al., 2008. Proteomics analysis unravels the functional repertoire of coronavirus nonstructural protein 3. J. Virol. 82, 5279–5294. https://doi.org/10.1128/jvi.02631-07.

Nick Pace, C., Martin Scholtz, J., 1998. A helix propensity scale based on experimental studies of peptides and proteins. Biophys. J. 75, 422–427. https://doi.org/10.1016/s0006-3495(98)77529-0.

Nielsen, R., 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. PLoS Biology. https://doi.org/10.1371/journal.pbio.0030170.

Oleary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., Mcveigh, R., et al., 2015. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44 https://doi.org/10.1093/nar/gkv1189.

Ong, E., Wong, M.U., Huffman, A., He, Y., 2020a. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. Front. Immunol. 11 https://doi.org/10.3389/fimmu.2020.01581.

Ong, E., Wong, M.U., Huffman, A., He, Y., 2020b. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. Front. Immunol. 11 https://doi.org/10.3389/fimmu.2020.01581.

Philip, S., Machado, J.P., Maldonado, E., Vasconcelos, V., Obrien, S.J., Johnson, W.E., Antunes, A., 2012. Fish lateral line innovation: insights into the evolutionary genomic dynamics of a unique mechanosensory organ. Mol. Biol. Evol. 29, 3887–3898. https://doi.org/10.1093/molbev/mss194.

Pillay, T., 2020. Gene of the month: the 2019-nCoV/SARS-CoV-2 novel coronavirus spike protein. J. Clin. Pathol. 73, 366–369. https://doi.org/10.1136/jclinpath-2020-206658.

Raamsman, MJ, Locker, JK, de Hooge, A, de Vries, AA, Griffiths, G, Vennema, H, Rottier, PJ., et al., 2000. Characterization of the coronavirus mouse hepatitis virus strain A59 small membrane protein E. J. Virol. (5), 2333–2342. https://doi.org/10.1128/jvi.74.5.2333-2342.2000.

Rambaut, A., Holmes, E., Hill, V., O'Toole, Á., McCrone, J., Ruis, C., Plessis, L., and Pybus, O. (2020). A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. bioRxiv preprint. doi: https://doi.org/10.1101/2020.04.17.046086.

Schoeman, D., Fielding, B., et al., 2019. Coronavirus envelope protein: current knowledge. Virol. J. 16 (69). https://doi.org/10.1186/s12985-019-1182-0.

Schubert, K., Karousis, E., Jomaa, A., Scaiola, A., Echeverria, B., Gurzeler, L., Leibundgut, M., Thiel, V., Mühlemann, O., and Ban, N. (2020). SARS-CoV-2 Nsp1 binds ribosomal mRNA channel to inhibit translation. doi:10.1101/2020.07.07.191676.

Shi, J., Wen, Z., Zhong, G., Yang, H., Wang, C., Huang, B., et al., 2020. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. Science. https://doi.org/10.1126/science.abb7015.

Shi, P., Su, Y., Li, R., Liang, Z., Dong, S., Huang, J., 2019. PEDV nsp16 negatively regulates innate immunity to promote viral proliferation. Virus Res. 265, 57–66. https://doi.org/10.1016/j.virusres.2019.03.005.

Song, Z., Xu, Y., Bao, L., Zhang, L., Yu, P., Qu, Y., et al., 2019. From SARS to MERS, thrusting coronaviruses into the spotlight. Viruses 11, 59. https://doi.org/10.3390/v11010059.

Simmonds, P., et al., 2020. Rampant C >U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short-and long-term evolutionary trajectories. Msphere 2020, 5. https://doi.org/10.1128/mSphere.00408-20.

Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577. https://doi.org/10.1080/10635150701472164.

Tang, Q., Song, Y., Shi, M., Cheng, Y., Zhang, W., Xia, X.-Q., 2015. Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. Sci. Rep. 5 https://doi.org/10.1038/srep17155.

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., et al., 2020. On the origin and continuing evolution of SARS-CoV-2. Natl. Sci. Rev. https://doi.org/10.1093/nsr/nwaa036.

Thoms, M., Buschauer, R., Ameismeier, M., Koepke, L., Denk, T., Hirschenberger, M., Kratzat, H., Hayn, M., Mackens-Kiani, T., Cheng, J., et al., 2020. Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. Science eabc8665. https://doi.org/10.1126/science.abc8665.

Torres, J, Parthasarathy, K, Lin, X, Saravanan, R, Kukol, A, Liu, DX., et al., 2006. Model of a putative pore: the pentameric alpha-helical bundle of SARS coronavirus E protein in lipid bilayers. Biophys. J. 91 (3), 938–947. https://doi.org/10.1529/biophysj.105.080119.

Walls, A., Park, Y., Tortorici, M., Wall, A., McGuire, A., Veesler, D, 2020. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 181, 281–292. https://doi.org/10.1016/j.cell.2020.02.058 e6.

Wan, Y., Shang, J., Graham, R., Baric, R., Li, F., 2020. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. J. Virol. 94 https://doi.org/10.1128/jvi.00127-20.

Wang, C., Horby, P.W., Hayden, F.G., Gao, G.F., 2020a. A novel coronavirus outbreak of global health concern. Lancet N. Am. Ed. 395, 470–473. https://doi.org/10.1016/s0140-6736(20)30185-9.

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., et al., 2020. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe 27, 325–328. https://doi.org/10.1016/j.chom.2020.02.001.

Wang, R., Hozumi, Y., Zheng, Y., Yin, C., Wei, G., et al., 2020b. Host immune response driving SARS-CoV-2 evolution. Viruses 12 (10). https://doi.org/10.3390/v12101095, 20201095.

Li, X., Giorgi, E., Marichannegowda, M., Gao, F., et al., 2020b. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Sci. Adv. 6 (27), eabb9153. https://doi.org/10.1126/sciadv.abb9153.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591. https://doi.org/10.1093/molbev/msm088.

Zhu, Z., Zhang, Z., Chen, W., Cai, Z., Ge, X., Zhu, H., et al., 2018. Predicting the receptor-binding domain usage of the coronavirus based on kmer frequency on spike protein. Infect. Genet. Evol. 61, 183–184. https://doi.org/10.1016/j.meegid.2018.03.028.