Mutational screening of the proteome of Sars-Cov-2 isolates: mutability of ORF3a, Nucleocapsid and Nsp2 proteins

Martina Bianchi[a], Domenico Benvenuto[b], Massimo Ciccozzi[b] and Stefano Pascarella[a*]

[a]Department of Biochemical sciences "A Rossi Fanelli", Sapienza University of Rome, 00185 Rome, Italy

[b]Unit of Medical Statistics and Molecular Epidemiology, University Campus Bio-Medico of Rome, Rome, Italy.

*Corresponding author:

Stefano Pascarella
Dipartimento di Scienze biochimiche "A. Rossi Fanelli"
Sapienza Università di Roma, 00185 Rome, Italy
e-mail: Stefano.Pascarella@uniroma1.it

Abstract

The Sars-CoV-2 is the causative agent of the current coronavirus disease pandemic. To effectively fight this pathogen, it is important to understand its evolution and the mechanism of adaptation to the host. A software workflow has been utilized to scan 26,016 Sars-CoV-2 genomes available in GISAID databank to analyse the distribution and frequency of mutations in the corresponding proteomes. A filtering procedure has been applied to remove data inconsistencies and redundancies. The number of observed mutations appears proportional to protein sequence length except for ORF3a, Nucleocapsid and Nsp2 that seem to accept more mutations than expected. The most pervasive mutations of the three proteins have been reported and the most variable and conservative regions mapped onto the respective sequences. The results suggest that these proteins may have a role in the adaptation of virus to new hosts and influence its pathogenicity and replication. These considerations prompt the experimental study and characterization of the three proteins.

Keywords: Sars-CoV-2; proteome; mutation; ORF3a; Nucleoprotein; Nsp2

1. Introduction

Coronavirus Disease (COVID-19) became almost suddenly, but not unexpectedly, a serious threat to human health [1–3]. The etiological agent of the disease is a Coronavirus classified as Sars-CoV-2 (Severe acute respiratory syndrome CoronaVirus 2) related to Sars-CoV. In 2002, Sars-CoV caused an outbreak of atypical and severe, often lethal, pneumonia in Guangdong province, China. This virus had a relatively low contagiousness and could be contained as to prevent worldwide spreading [4]. Coronaviruses are positive, single stranded RNA (+ssRNA) viruses that possess a genome of 27-32 kb in size. Sars-CoV-2 genome shares about 79% nucleotide sequence identity to Sars-CoV [5]. The genome codes for two large overlapping polyproteins that are processed by intracellular proteolysis into nonstructural proteins involved in virus replication and assembly [6,7]. A set of structural accessory proteins are also coded in the genome [8].

The coronaviruses are intrinsically promiscuous and can be hosted by several species. Indeed, it is now accepted that the current pandemic has been ignited by a cross-species virus transmission from Pangolin and/or Bat to humans, at Wuhan, China [2,9–11].

Like many viruses, the CoV evolves and adapts to the host through accumulation of synonymous and non-synonymous mutations [12] generated by several mechanisms including fidelity of RNA-dependent-RNA-polymerase [13]. Many initiatives are currently ongoing to develop effective diagnostic tools, vaccines and therapeutic strategies able to prevent and fight Sars-CoV-2 infections and diseases. In this context, it is important to understand the dynamics of evolution of the virus and to study how its proteome changes. Indeed, modification of specific virus proteins considered promising targets may put at risk the efficacy of drugs or vaccines. It is known that even single mutations in specific proteins can change pathogenicity or contagiousness of viruses [14,15].

In this work, a software workflow able to carry out a quick, systematic and repeatable screening of the Sars-CoV-2 genome isolates to scrutinize the frequency of specific mutations in each of the protein expressed by the virus, is reported. The workflow has been utilized to scan 26,016 Sars-CoV-2 genomes available in GISAID databank [16]. Three proteins, ORF3a, Nucleocapsid and Nsp2 seem to be more prone to mutations than expected. The most pervasive mutations of the three proteins are described and the most variable/conservative regions mapped onto the sequences.

2. Materials and Methods

The mutation screening of the proteins coded by different Sars-CoV-2 genome isolates has been carried out using the following computer workflow:

a. Sars-CoV-2 genome sequences have been downloaded in the standard FASTA format from GISAID repository at www.gisaid.org [16]. Since the quality of the sequences is not uniform, only complete sequences deposited with a high degree of coverage has been downloaded using the filters provided by the GISAID server.
b. The downloaded set of genomic sequences has been reformatted as a BLAST database using the "makeblastdb" tool in the BLAST suite [17].
c. Tblastn, a BLAST tool able to search a protein sequence against a translated nucleotide sequence database, has been applied to extract all the corresponding coding segments in each variant isolate from the local Sars-CoV-2 BLAST genomic database.
d. At the end of the process, each reference protein had a set of cognate variants. To avoid bias in statistics, all the retrieved sequences have been translated and further filtered: incomplete sequences or sequences containing ambiguous codons (resulting in undetermined residues) has been eliminated. This step relied on the tools available in the EMBOSS suite [18] along with Linux bash shell commands such as "grep", "awk" and "sed".
e. At this stage, the data are intrinsically redundant because many copies of the same protein identical to the reference will be included in the data set while other copies will contain single or multiple mutations. The clustering algorithm implemented in "cd-hit" [19] has been used to group all the identical sequences linked to the same reference. At the end of the procedure, a set of clusters was obtained each one representing a variant of the reference protein. From each cluster, only one representing sequence has been considered and further processed.
f. All the representative variants have been multiply aligned to the reference protein with the program MAFFT [20].
g. A script has been written in R language under the Rstudio environment for statistical analyses and graphical output. The R script utilized input and output functions from the bio3d package [21]. Plots have been drawn with the ggplot2 package.

Transmembrane and secondary structure prediction utilized TMHMM [22], PsiPred [23] and Jpred [24]. Multiple sequence alignments display and editing relied on Jalview [25].

3. Results

From the GISAID repository accessed on June 2020, 26,016 genome isolates were collected. RefSeq [26] sequences used as the reference Sars-CoV-2 proteins in this analysis are listed in **Table 1**. The complete list is reported in **Supplementary File S1**. Each one of these sequences has been used as a Tblastn query to delimit the corresponding coding segment in every Sars-CoV-2 genome. **Table 1** reports the number of sequences remaining after filtering and removing the sequences containing undetermined residues or stop codons. Results of the "cd-hit" clustering applied to each set of reference proteins is also reported in **Table 1**. For each reference protein, identical variants have been assigned to the same cluster from which a representative sequence has been selected by "cd-hit". The number of sequences belonging to each cluster has been defined "cluster size". The cluster size should approximate the pervasiveness of its variant.

Fig. 1 reports number of variant clusters compared with the sequence length of the corresponding reference protein. The distribution suggests that the number of variants observed for each protein, reflecting the number of different sequence mutations, is directly proportional to the sequence length (correlation coefficient = 0.93). However, there are three notable exceptions: ORF3a, Nucleocapsid and Nsp2 which appears to be hit by an amount of mutations significantly higher than expected from their length.

Variants of the three outliers have been studied in detail. Cluster size distribution, defined as the number of sequences belonging to each cluster of the same reference protein, has been calculated.. In all the cases, the most populated cluster corresponds to the reference sequence. Details on the mutations characterizing the clusters of each sequence are reported in **Table 2**. The cluster size distribution suggests that only the second peak is significantly populated and represents more than 10% of the cluster population. The clusters of Orf3a are characterized by single sequence changes while Nucleocapsid and Nsp2 have one variant with two simultaneous mutations. Moreover, Nsp2 variant number 4 is characterized by the deletion of the residue Asp268.

3.1 ORF3a

Fig. 2 reports the multiple sequence alignment among the reference sequence and the four most frequent variants. ORF3a is an integral membrane protein possessing three transmembrane helices located at the N-terminal side of the polypeptide chain [27,28]. The most frequent variant is represented by the substitution Q → H in position 57 of the multiple sequence alignment, occurring at the endoplasmic reticulum (ER) edge of the first transmembrane helix. This substitution conserves the hydrophilic character of the site although His has a side chain dissociable proton lacking in Gln. The second most frequent variant is G → V in position 251 (**Fig. 2**). This position is within the intra-ER portion of the chain and it is at the C-terminal side of a predicted β-strand. The site is also predicted to be exposed to the solvent. This substitution should change the local flexibility of the backbone while increasing hydrophobicity. The other two substitutions are G → V in position 196 and V → L in position 13 (**Fig. 2**). The former substitution is in a loop between two predicted β-strands in the intra-ER portion, exposed to the solvent. The replacement of a Gly residue with a Val should alter local backbone flexibility and increase hydrophobicity. At variance with the other mutations, V → L in position 13 is in the intra-cytoplasmic portion of the chain, before the first transmembrane helix. The position is predicted to be partly buried within the polypeptide chain. Apparently, the substitution does not have significant effect on the local hydrophobicity. Overall, the most conserved portion of the protein appears to be in the ER domain (**Supplementary Fig. S1**).

3.2 Nucleocapsid

Fig. 3 reports the multiple sequence alignment among the reference sequence and the four most frequent variants. Nucleocapsid protein has been extensively studied in Sars-CoV and it has been demonstrated to be composed of two domains connected by a linker [29]. The N-terminal portion is the RNA binding domain while the C-terminal part implements the oligomerization function. Sars-CoV-2 nucleoprotein has a similar organization. The crystallographic structures of the N- and C-terminal portions of the Sars-CoV-2 nucleoprotein are available in the PDB. The structure identified by the PDB code 6Yi3 covers the nucleoprotein N-terminal segment 43-180, while 6WZQ is the C-terminal portion encompassed by sequence positions 247-364. The most frequent variant of this protein is characterized by the dyad KR replacing RG at positions 203-204 within the linker region. The replacement of a Gly with an Arg introduces a positively charged residue often involved in the interaction with nucleic acids in nucleic acid binding proteins. This region is predicted to be exposed and in loop conformation (**Fig. 3**). The mutations characterizing the other variants introduce a Leu residue in place or Ser or Pro. Ser is replaced at positions 194 and 197 close to the N-terminal domain in a region predicted to be a loop. The substituted Pro is found at position 13 in the N-terminal segment of the sequence. This region also is predicted to be in a loop. Overall, the most conserved regions of this protein correspond to the two functional domains (**Supplementary Fig. S2**).

3.3 Nsp2

Fig. 4 displays the multiple sequence alignment among the reference sequence and the four most frequent variants of the Nsp2 protein. The most frequent mutation is T → I in position 85 at a predicted exposed site within a loop. The other most frequent variant possesses two mutations located in the C-terminal domain of the Nsp2 sequence. The mutations are I → V in position 559 at the beginning of a predicted β-strand and P → S at position 585 in a loop. The other two variants have a deletion of Asp at position 268 and G → D substitution at position 212, respectively, both at the beginning of predicted α-helices. Overall, the conserved regions appear homogeneously distributed over the sequence (**Supplementary Fig. S3**).

4. Discussion

Sars-CoV-2 virus is seriously threatening global health and it is claiming many lives. To effectively fight this pathogen, it is important to understand its evolution and the mechanism of adaptation to the host. This information will also contribute to create a knowledge-base to face the future epidemics of zoonotic origin. In this work, the proteomes of the genome isolates of Sars-CoV-2 virus deposited in the repository GISAID have been scrutinized for mutations. For each protein coded by the virus genome, all the variants available, namely all the unique mutant forms, have been collected with the aid of a computer workflow. Proteins are the effectors of virus biological functions and endure the selective pressure due to the process of adaptation to various hosts and environmental conditions [13,30]. The number of different variants observed for each protein is proportional to the sequence length: this may suggest that each residue has an approximately similar probability of being hit by a mutation. This is apparently not true for the three proteins ORF3a, Nucleocapsid and Nsp2 where a fraction of mutations higher than expected has been observed. This fact prompts for a detailed consideration of the three proteins and of their role in virus pathogenesis.

ORF3a is an integral membrane and in Sars-CoV it has been demonstrated that it activates the NLRP3 inflammasome and it is a potent stimulator of pro–IL-1β gene transcription [27]. A recent study suggested that the Sars-CoV-2 protein is linked to virulence, infectivity, ion channel formation, and virus release [28]. A link between the mutability of this protein and the process of virus adaptation to the new human host can be hypothesized. The mutated sites are in the intra ER region and may be subject to functional constraints reflected by the local sequence conservation. However, it should be noted that the C-terminal segment is rather variable (**Supplementary Fig. S1**).

Nucleoprotein has been studied and characterized in Sars-CoV [29,31]. Several crystallographic structures from Sars-Cov and Sars-CoV-2 are now deposited and available in the PDB. The major function of this protein is the packaging of the viral genome into a helical ribonucleocapsid. Therefore, it plays a fundamental role during viral self-assembly. However, nucleoprotein is also a multifaceted protein. Indeed, the SARS-CoV nucleoprotein has been shown to modulate the host cellular machinery and to play regulatory roles during viral life cycle [32,33]. The map of distribution of mutations along the sequence (**Supplementary Fig. S2**) suggests that the N- and C-terminal domains tend to be more conserved while the linker region is more receptive to mutations. The N- and C-terminal extremities also appear more prone to mutate. Linkers are often considered passive modules within a multidomain protein. However, it has been suggested that they may play important roles [34–37]. In the case of Nucleocapsid protein, the most frequent mutations change the physico-chemical properties of the sites except for the substitution R203 → K. This may suggest that the substitution has an impact on the role of the linker region in virus assembly and pathogenesis.

The role of Nsp2 is still elusive. It has been suggested [38] that the Sars and Mers Nsp2 is not essential for virus replication. However, an interaction with two host proteins, prohibitin 1 (PHB1) and PHB2 has been identified. This may indicate that Nsp2 is involved in the alteration of intracellular host signalling during SARS-CoV infections [38,39]. Overall the distribution of mutated sites within the sequence appears rather homogeneous (**Supplementary Fig. S3**) even though the C-terminal portion is highly variable. The most frequent mutations change the physico-chemical properties of the sites in which they occur. Nsp2 displays also a variant possessing a deletion at the beginning of an α-helix in position 268 which has been initially observed in Europe [40].

The striking ability of viruses to adapt to new hosts and environments is reflected by their capacity to generate variations in a short time. In general, ssRNA viruses mutate faster than other viruses. In this case, three Sars-CoV-2 proteins appear to have accumulated a number of mutations higher than expected. The prevalence of these mutations in the genome isolates suggests also that the observed changes are related to the process of virus adaptation to the host possibly conferring advantages in terms of protein stability and/or replication efficiency. These considerations prompt for further experimental investigations to test the role of these proteins in virus pathogenicity. Finally, the workflow here described can be easily utilized to constantly monitor the evolution and modification of this and other viruses.

Acknowledgments

This work has been in part funded by a grant to SP from Sapienza University of Rome [grant number RP11916B74B27C4D].

References

[1] Y.Z. Zhang, E.C. Holmes, A Genomic Perspective on the Origin and Emergence of SARS-CoV-2, Cell. 181 (2020) 223–227. https://doi.org/10.1016/j.cell.2020.03.035.

[2] D. Benvenuto, M. Giovanetti, M. Salemi, M. Prosperi, C. De Flora, L.C. Junior Alcantara, S. Angeletti, M. Ciccozzi, The global spread of 2019-nCoV: a molecular evolutionary analysis., Pathog. Glob. Health. (2020) 1–4. https://doi.org/10.1080/20477724.2020.1725339.

[3] M. Ciotti, S. Angeletti, M. Minieri, M. Giovannetti, D. Benvenuto, S. Pascarella, C. Sagnelli, M. Bianchi, S. Bernardini, M. Ciccozzi, COVID-19 Outbreak: An Overview, Chemotherapy. (2020). https://doi.org/10.1159/000507423.

[4] J.S.M. Peiris, Y. Guan, K.Y. Yuen, Severe acute respiratory syndrome, Nat. Med. 10 (2004) S88–S97. https://doi.org/10.1038/nm1143.

[5] D. Kaul, An overview of coronaviruses including the SARS-2 coronavirus - Molecular biology, epidemiology and clinical implications, Curr. Med. Res. Pract. 10 (2020) 54–64. https://doi.org/10.1016/j.cmrp.2020.04.001.

[6] Y. Qiu, K. Xu, Functional studies of the coronavirus nonstructural proteins, STEMedicine. 1 (2020). https://doi.org/10.37175/stemedicine.v1i2.39.

[7] M. Bartlam, H. Yang, Z. Rao, Structural insights into SARS coronavirus proteins, Curr. Opin. Struct. Biol. 15 (2005) 664–672. https://doi.org/10.1016/j.sbi.2005.10.004.

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

[9] D. Benvenuto, M. Giovanetti, A. Ciccozzi, S. Spoto, S. Angeletti, M. Ciccozzi, The 2019-new coronavirus epidemic: Evidence for virus evolution., J. Med. Virol. 92 (2020) 455–459. https://doi.org/10.1002/jmv.25688.

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

[11] M. Bianchi, D. Benvenuto, M. Giovanetti, S. Angeletti, M. Ciccozzi, S. Pascarella, Sars-CoV-2 Envelope and Membrane Proteins: Structural Differences Linked to Virus Characteristics?, Biomed Res. Int. 2020 (2020) 1–6. https://doi.org/10.1155/2020/4389089.

[12] T. Phan, Genetic diversity and evolution of SARS-CoV-2, Infect. Genet. Evol. 81 (2020) 104260. https://doi.org/10.1016/j.meegid.2020.104260.

[13] R. Sanjuán, P. Domingo-Calap, Mechanisms of viral mutation, Cell. Mol. Life Sci. 73 (2016) 4433–4448. https://doi.org/10.1007/s00018-016-2299-6.

[14] Y. Sakai, K. Kawachi, Y. Terada, H. Omori, Y. Matsuura, W. Kamitani, Two-amino acids change in the nsp4 of SARS coronavirus abolishes viral replication., Virology. 510 (2017) 165–174. https://doi.org/10.1016/j.virol.2017.07.019.

[15] C.B. Hwang, K.L. Ruffner, D.M. Coen, A point mutation within a distinct conserved region of the herpes simplex virus DNA polymerase gene confers drug resistance., J. Virol. 66 (1992) 1774–1776. https://doi.org/10.1128/jvi.66.3.1774-1776.1992.

[16] Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality., Euro Surveill. Bull. Eur. Sur Les Mal. Transm. = Eur. Commun. Dis. Bull. 22 (2017). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

[17] S. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

[18] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite., Trends Genet. 16 (2000) 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.

[19] W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics. 22 (2006) 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

[20] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability., Mol. Biol. Evol. 30 (2013) 772–780. https://doi.org/10.1093/molbev/mst010.

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

[22] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes., J. Mol. Biol. 305 (2001) 567–580. https://doi.org/10.1006/jmbi.2000.4315.

[23] L.J. McGuffin, K. Bryson, D.T. Jones, The PSIPRED protein structure prediction server., Bioinformatics. 16 (2000) 404–405. https://doi.org/10.1093/bioinformatics/16.4.404.

[24] A. Drozdetskiy, C. Cole, J. Procter, G.J. Barton, JPred4: A protein secondary structure prediction server, Nucleic Acids Res. 43 (2015) W389–W394. https://doi.org/10.1093/nar/gkv332.

[25] A.M. Waterhouse, J.B. Procter, D.M.A. Martin, M. Clamp, G.J. Barton, Jalview Version 2-A multiple sequence alignment editor and analysis workbench, Bioinformatics. 25 (2009) 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

[26] K.D. Pruitt, T. Tatusova, G.R. Brown, D.R. Maglott, NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy, Nucleic Acids Res. 40 (2012) D130–D135. https://doi.org/10.1093/nar/gkr1079.

[27] K.-L. Siu, K.-S. Yuen, C. Castaño-Rodriguez, Z.-W. Ye, M.-L. Yeung, S.-Y. Fung, S. Yuan, C.-P. Chan, K.-Y. Yuen, L. Enjuanes, D.-Y. Jin, Severe acute respiratory syndrome coronavirus

ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC, FASEB J. 33 (2019) 8865–8877. https://doi.org/10.1096/fj.201802418R.

[28] E. Issa, G. Merhi, B. Panossian, T. Salloum, S. Tokajian, SARS-CoV-2 and ORF3a: Nonsynonymous Mutations, Functional Domains, and Viral Pathogenesis, MSystems. 5 (2020) e00266-20. https://doi.org/10.1128/mSystems.00266-20.

[29] C. Chang, M.-H. Hou, C.-F. Chang, C.-D. Hsiao, T. Huang, The SARS coronavirus nucleocapsid protein--forms and functions., Antiviral Res. 103 (2014) 39–50. https://doi.org/10.1016/j.antiviral.2013.12.009.

[30] D. Benvenuto, S. Angeletti, M. Giovanetti, M. Bianchi, S. Pascarella, R. Cauda, M. Ciccozzi, A. Cassone, Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy, J. Infect. 81 (2020). https://doi.org/10.1016/j.jinf.2020.03.058.

[31] Y. Hu, W. Li, T. Gao, Y. Cui, Y. Jin, P. Li, Q. Ma, X. Liu, C. Cao, The Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Inhibits Type I Interferon Production by Interfering with TRIM25-Mediated RIG-I Ubiquitination, J. Virol. 91 (2017) e02143-16. https://doi.org/10.1128/JVI.02143-16.

[32] M. Surjit, B. Liu, V.T.K. Chow, S.K. Lal, The nucleocapsid protein of severe acute respiratory syndrome-coronavirus inhibits the activity of cyclin-cyclin-dependent kinase complex and blocks S phase progression in mammalian cells., J. Biol. Chem. 281 (2006) 10669–10681. https://doi.org/10.1074/jbc.M509233200.

[33] M. Surjit, S.K. Lal, The SARS-CoV nucleocapsid protein: a protein with multifarious activities., Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. 8 (2008) 397–405. https://doi.org/10.1016/j.meegid.2007.07.004.

[34] T. Milano, S. Angelaccio, A. Tramonti, M.L. Di Salvo, R. Contestabile, S. Pascarella, Structural properties of the linkers connecting the N- and C- terminal domains in the MocR bacterial transcriptional regulators, Biochim. Open. 3 (2016) 8–18. https://doi.org/10.1016/j.biopen.2016.07.002.

[35] T. Milano, A. Gulzar, D. Narzi, L. Guidoni, S. Pascarella, Molecular dynamics simulation unveils the conformational flexibility of the interdomain linker in the bacterial transcriptional regulator GabR from Bacillus subtilis bound to pyridoxal 5'-phosphate, PLoS One. 12 (2017) e0189270. https://doi.org/10.1371/journal.pone.0189270.

[36] R.A. George, J. Heringa, An analysis of protein domain linkers: their classification and role in protein folding, Protein Eng. Des. Sel. 15 (2002) 871–879. https://doi.org/10.1093/protein/15.11.871.

[37] D. Luo, N. Wei, D.N. Doan, P.N. Paradkar, Y. Chong, A.D. Davidson, M. Kotaka, J. Lescar, S.G. Vasudevan, Flexibility between the protease and helicase domains of the dengue virus NS3 protein conferred by the linker region and its functional implications, J. Biol. Chem. 285 (2010) 18817–18827. https://doi.org/10.1074/jbc.M109.090936.

[38] R.L. Graham, A.C. Sims, R.S. Baric, M.R. Denison, The nsp2 proteins of mouse hepatitis virus and SARS coronavirus are dispensable for viral replication., Adv. Exp. Med. Biol. 581 (2006) 67–72. https://doi.org/10.1007/978-0-387-33012-9_10.

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

[40] A. Bal, G. Destras, A. Gaymard, M. Bouscambert-Duchamp, M. Valette, V. Escuret, E. Frobert, G. Billaud, S. Trouillet-Assant, V. Cheynet, K. Brengel-Pesce, F. Morfin, B. Lina, L. Josset, Molecular characterization of SARS-CoV-2 in the first COVID-19 cluster in France

reveals an amino acid deletion in nsp2 (Asp268del), Clin. Microbiol. Infect. 26 (2020) 960. https://doi.org/10.1016/j.cmi.2020.03.020.

Table 1

Data set utilized

Reference RefSeq code	Protein denomination	No. of GISAID sequences[a]	No. of variants[b]	Sequence length
Yp_009724390	Spike glycoprotein	21,321	864	1273
Yp_009724391	ORF3a	25,381	488	275
Yp_009724392	Envelope protein	25,870	55	83
Yp_009724393	Membrane glycoprotein	25,664	127	222
Yp_009724394	ORF6	25,861	50	61
Yp_009724395	ORF7a	25,195	130	121
Yp_009724396	ORF8	25,750	143	121
Yp_009724397	Nucleocapsid	25,210	606	419
Yp_009725255	ORF10	25,767	39	38
Yp_009725296	ORF7b	25,467	45	43
Yp_009725297	Leader	25,585	168	180
Yp_009725298	Nsp2	24,949	637	638
Yp_009725299	Nsp3	23,722	1199	1945
Yp_009725300	Nsp4	25,253	266	500
Yp_009725301	3C-like protease	25,692	157	306
Yp_009725302	Nsp6	25,461	182	290
Yp_009725303	Nsp7	25,946	47	83
Yp_009725304	Nsp8	25,811	95	198
Yp_009725305	Nsp9	25,890	65	113
Yp_009725306	Nsp10	25,639	55	139
Yp_009725307	RdRp	25,143	504	932
Yp_009725308	Helicase	25,123	322	601
Yp_009725309	3’-to-5’ exonuclease	21,302	264	527
Yp_009725310	endoRNAse	25,142	222	346
Yp_009725311	2’-O-ribose MT	25,113	156	298
Yp_009725312	Nsp11	25,725	8	13

[a] Number of sequences after filtering

[b] Number of clusters found by cd-hit

Table 2

Top four most frequent variants of ORF3a, Nucleocapsid and Nsp2

Protein denomination	Cluster[a]	Alignment position[b]	Mutations[c]	Cluster fractional size[d]	Representative isolate (EPI_ISL)[e]
ORF3a	ref			58.5	458798
	2	57	Q →H	24.9	456153
	3	252	G →V	7.0	458794
	4	13	V→L	2.0	444164
	5	196	G→V	1.2	422803
Nucleocapsid	ref			63.3	458797
	2	203,204	R→K, G→R	21.5	458798
	3	194	S→L	1.7	444175
	4	13	P→L	1.3	456197
	5	197	S→L	1.2	422803
Nsp2	ref			65.3	429802
	2	85	T→I	18.3	424238
	3	559,585	I→V, P→S	2.4	444153
	4	268	D → *Δ*	2.2	430016
	5	212	G→D	1.4	417412

[a] "ref" indicates the cluster corresponding to the reference protein

[b] positions refer to the alignments in Figures 2, 3 and 4

[c] "*Δ*" indicates deletion

[d] Fractional cluster size relative to the entire data set

[e] GISAID genome codes

Figure legends

Figure 1

Distribution of cluster size versus sequence length for each protein. The linear regression fit is displayed. Grey band around the line marks the 95% confidence interval. Outliers are labelled according to corresponding protein denotation.

Figure 2

Alignment of the representative sequences of the top five populated clusters of ORF3a protein. Sequence are labelled with RefSeq or GISAID codes. Column colour intensity is proportional to the sequence conservation. Lines labelled with Jpred, TM2 and TM report the secondary structure prediction according to Jpred, and the transmembrane helix predictions from TMHMM and Psipred, respectively. Red bars denote helices while blue arrows are β-strands. Grey line are loop regions. Blu and orange boxes denote the segments predicted to be exposed in the cytoplasm (Cyt) and in the ER, respectively.

Figure 3

Alignment of the representative sequences of the top five populated clusters of Nucleocapsid protein. Sequence are labelled with RefSeq or GISAID codes. Column colour intensity is proportional to the sequence conservation. Line labelled with Jpred, reports the secondary structure prediction according to Jpred except within the transparent boxes. Red bars denote helices while blue arrows are β-strands. Grey line are loop regions. Transparent grey and red boxes mark the portions of the protein corresponding to the RNA binding and dimerization domains, respectively, solved by X-ray crystallography. Secondary structures within these regions were derived from the crystallographic data.

Figure 4

Alignment of the representative sequences of the top five populated clusters of Nsp2 protein. Sequence are labelled with RefSeq or GISAID codes. Column colour intensity is proportional to the sequence conservation. Line labelled with Jpred, reports the secondary structure prediction according to Jpred. Red bars denote helices while blue arrows are β-strands. Grey line are loop regions.

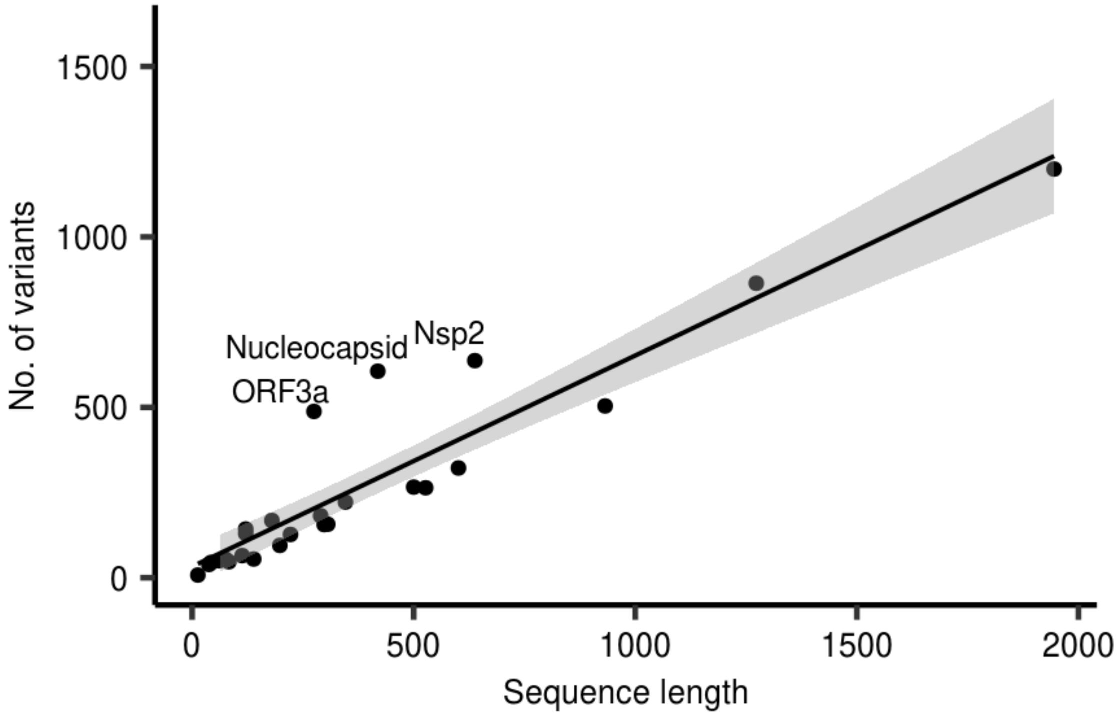
No. of variants
1500
1000
500
0
Sequence length
0
500
1000
1500
2000
Nucleocapsid
Nsp2
ORF3a

Figure 1

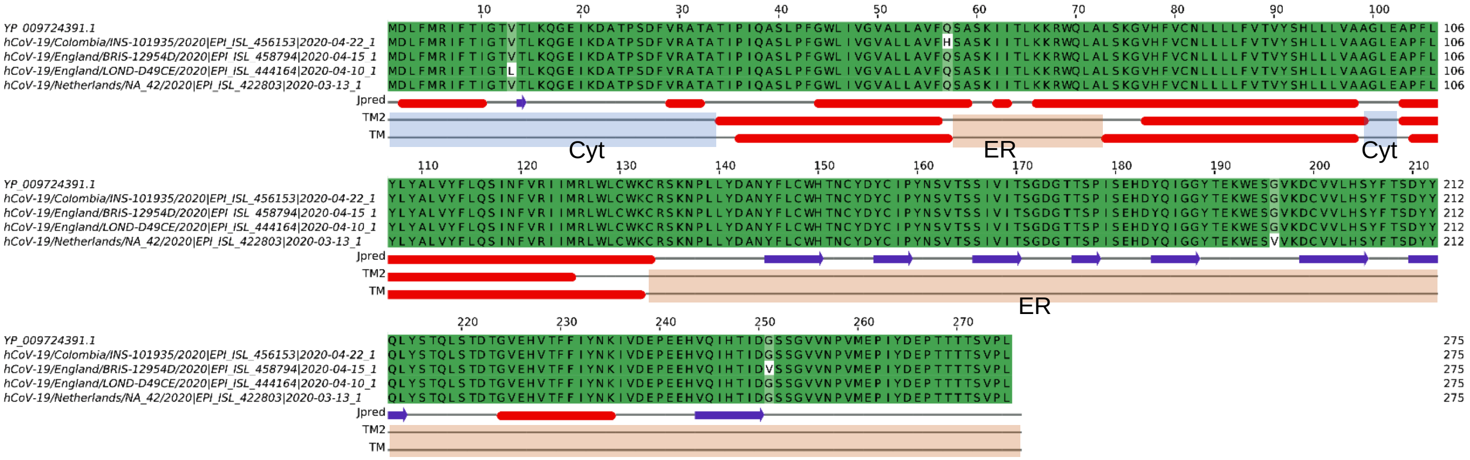
10
20
30
40
50
60
70
80
90
100
YP_009724391.1 MDLFMRIFTIGTVTLKQGEIKDATPSDFVRATATIPIQASLPFGWLIVGVALLAVFQSASKIITLKKRWQLALSKGVHFVCNLLLLFVTVYSHLLLVAAGLEAPFL 106
hCoV-19/Colombia/INS-101935/2020|EPI_ISL_456153|2020-04-22_1 MDLFMRIFTIGTVTLKQGEIKDATPSDFVRATATIPIQASLPFGWLIVGVALLAVFHSASKIITLKKRWQLALSKGVHFVCNLLLLFVTVYSHLLLVAAGLEAPFL 106
hCoV-19/England/BRIS-12954D/2020|EPI_ISL_458794|2020-04-15_1 MDLFMRIFTIGTVTLKQGEIKDATPSDFVRATATIPIQASLPFGWLIVGVALLAVFQSASKIITLKKRWQLALSKGVHFVCNLLLLFVTVYSHLLLVAAGLEAPFL 106
hCoV-19/England/LOND-D49CE/2020|EPI_ISL_444164|2020-04-10_1 MDLFMRIFTIGTLTLKQGEIKDATPSDFVRATATIPIQASLPFGWLIVGVALLAVFQSASKIITLKKRWQLALSKGVHFVCNLLLLFVTVYSHLLLVAAGLEAPFL 106
hCoV-19/Netherlands/NA_42/2020|EPI_ISL_422803|2020-03-13_1 MDLFMRIFTIGTVTLKQGEIKDATPSDFVRATATIPIQASLPFGWLIVGVALLAVFQSASKIITLKKRWQLALSKGVHFVCNLLLLFVTVYSHLLLVAAGLEAPFL 106
Jpred
TM2
TM
Cyt
ER
Cyt
110
120
130
140
150
160
170
180
190
200
210
YP_009724391.1 YLYALVYFLQSINFVRIIMRLWLCWKCRSKNPLLYDANYFLCWHTNCYDCIPYNSVTSSIVITSGDGTTSPISEHDYQIGGYTEKWESGVKDCVVLHSYFTSDYY 212
hCoV-19/Colombia/INS-101935/2020|EPI_ISL_456153|2020-04-22_1 YLYALVYFLQSINFVRIIMRLWLCWKCRSKNPLLYDANYFLCWHTNCYDCIPYNSVTSSIVITSGDGTTSPISEHDYQIGGYTEKWESGVKDCVVLHSYFTSDYY 212
hCoV-19/England/BRIS-12954D/2020|EPI_ISL_458794|2020-04-15_1 YLYALVYFLQSINFVRIIMRLWLCWKCRSKNPLLYDANYFLCWHTNCYDCIPYNSVTSSIVITSGDGTTSPISEHDYQIGGYTEKWESGVKDCVVLHSYFTSDYY 212
hCoV-19/England/LOND-D49CE/2020|EPI_ISL_444164|2020-04-10_1 YLYALVYFLQSINFVRIIMRLWLCWKCRSKNPLLYDANYFLCWHTNCYDCIPYNSVTSSIVITSGDGTTSPISEHDYQIGGYTEKWESGVKDCVVLHSYFTSDYY 212
hCoV-19/Netherlands/NA_42/2020|EPI_ISL_422803|2020-03-13_1 YLYALVYFLQSINFVRIIMRLWLCWKCRSKNPLLYDANYFLCWHTNCYDCIPYNSVTSSIVITSGDGTTSPISEHDYQIGGYTEKWESVVKDCVVLHSYFTSDYY 212
Jpred
TM2
TM
ER
220
230
240
250
260
270
YP_009724391.1 QLYSTQLSTDTGVEHVTFFIYNKIVDEPEEHVQIHTIDGSSGVVNPVMEPIYDEPTTTTSVPL 275
hCoV-19/Colombia/INS-101935/2020|EPI_ISL_456153|2020-04-22_1 QLYSTQLSTDTGVEHVTFFIYNKIVDEPEEHVQIHTIDGSSGVVNPVMEPIYDEPTTTTSVPL 275
hCoV-19/England/BRIS-12954D/2020|EPI_ISL_458794|2020-04-15_1 QLYSTQLSTDTGVEHVTFFIYNKIVDEPEEHVQIHTIDVSSGVVNPVMEPIYDEPTTTTSVPL 275
hCoV-19/England/LOND-D49CE/2020|EPI_ISL_444164|2020-04-10_1 QLYSTQLSTDTGVEHVTFFIYNKIVDEPEEHVQIHTIDGSSGVVNPVMEPIYDEPTTTTSVPL 275
hCoV-19/Netherlands/NA_42/2020|EPI_ISL_422803|2020-03-13_1 QLYSTQLSTDTGVEHVTFFIYNKIVDEPEEHVQIHTIDGSSGVVNPVMEPIYDEPTTTTSVPL 275
Jpred
TM2
TM

Figure 2

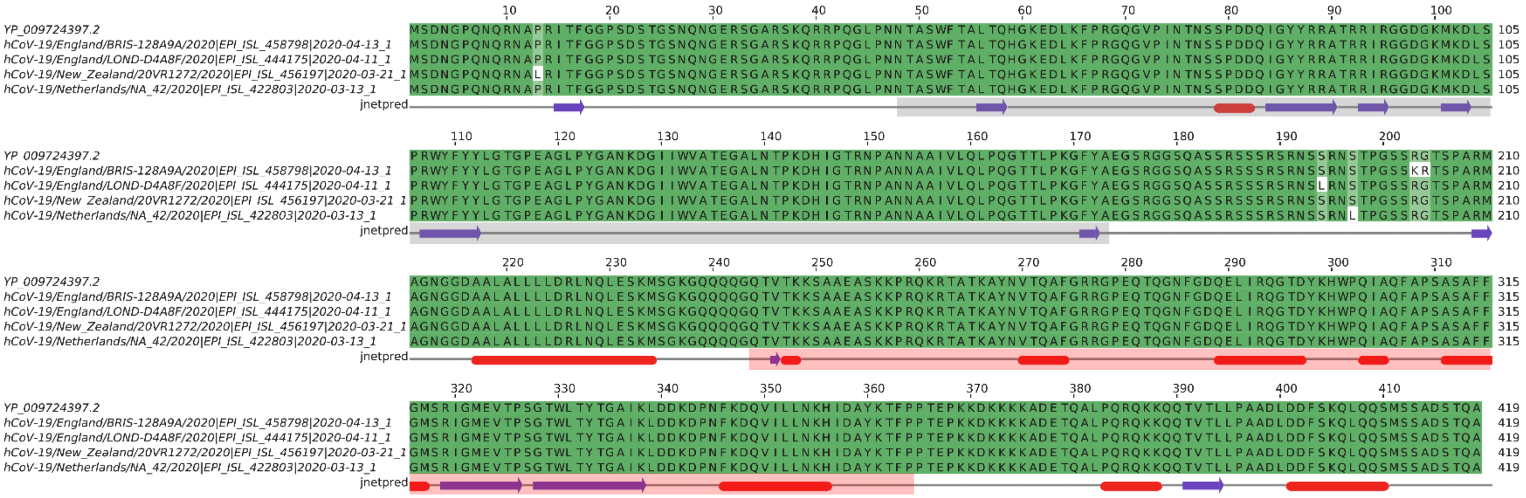
YP_009724397.2
hCoV-19/England/BRIS-128A9A/2020|EPI_ISL_458798|2020-04-13_1
hCoV-19/England/LOND-D4A8F/2020|EPI_ISL_444175|2020-04-11_1
hCoV-19/New_Zealand/20VR1272/2020|EPI_ISL_456197|2020-03-21_1
hCoV-19/Netherlands/NA_42/2020|EPI_ISL_422803|2020-03-13_1
jnetpred
10
20
30
40
50
60
70
80
90
100
105
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
315
320
330
340
350
360
370
380
390
400
410
419

Figure 3

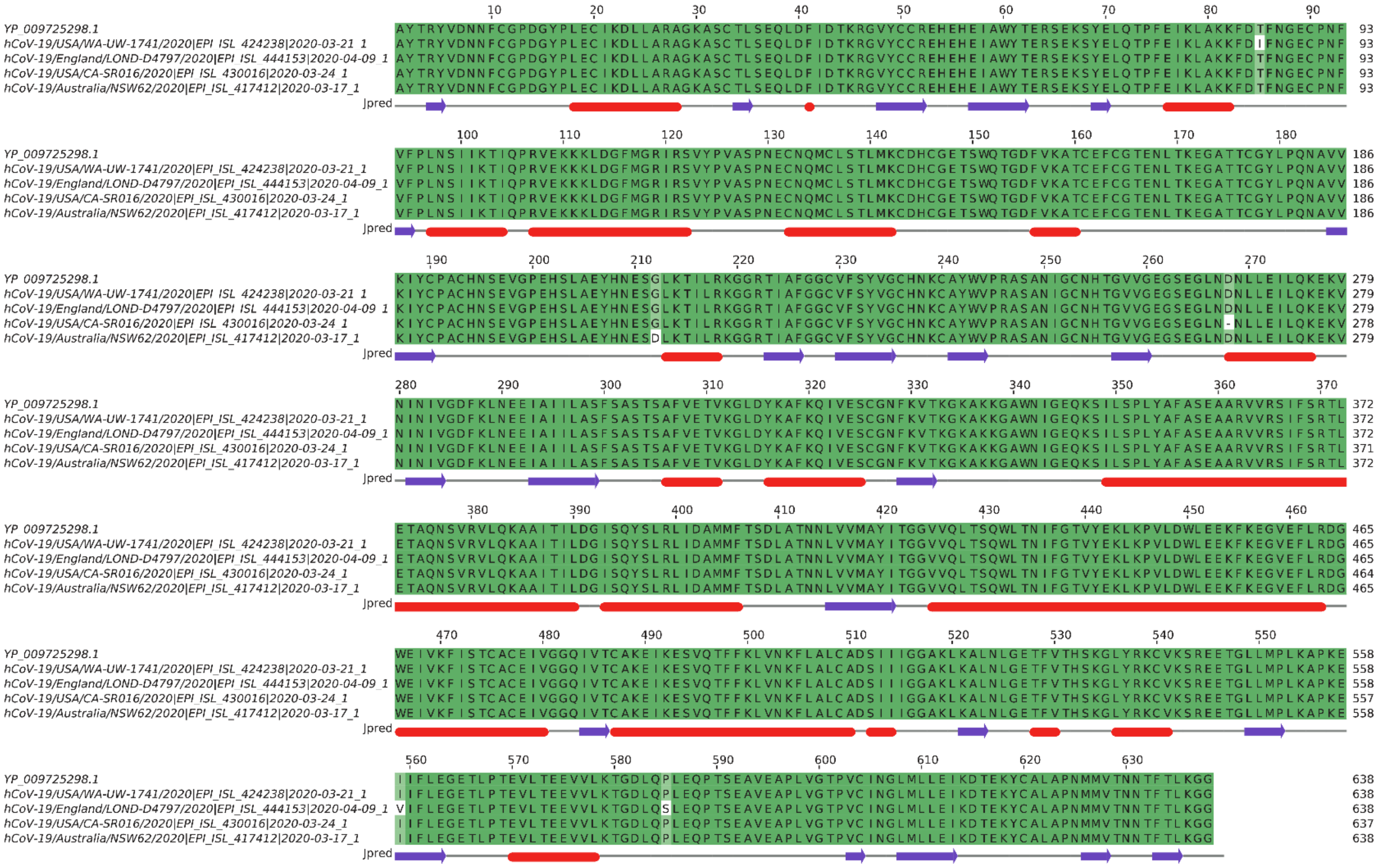
10 20 30 40 50 60 70 80 90
YP_009725298.1 AYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNF 93
hCoV-19/USA/WA-UW-1741/2020|EPI_ISL_424238|2020-03-21_1 AYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDIFNGECPNF 93
hCoV-19/England/LOND-D4797/2020|EPI_ISL_444153|2020-04-09_1 AYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNF 93
hCoV-19/USA/CA-SR016/2020|EPI_ISL_430016|2020-03-24_1 AYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNF 93
hCoV-19/Australia/NSW62/2020|EPI_ISL_417412|2020-03-17_1 AYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNF 93
Jpred
100 110 120 130 140 150 160 170 180
YP_009725298.1 VFPLNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTGDFVKATCEFCGTENLTKEGATTCGYLPQNAVV 186
hCoV-19/USA/WA-UW-1741/2020|EPI_ISL_424238|2020-03-21_1 VFPLNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTGDFVKATCEFCGTENLTKEGATTCGYLPQNAVV 186
hCoV-19/England/LOND-D4797/2020|EPI_ISL_444153|2020-04-09_1 VFPLNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTGDFVKATCEFCGTENLTKEGATTCGYLPQNAVV 186
hCoV-19/USA/CA-SR016/2020|EPI_ISL_430016|2020-03-24_1 VFPLNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTGDFVKATCEFCGTENLTKEGATTCGYLPQNAVV 186
hCoV-19/Australia/NSW62/2020|EPI_ISL_417412|2020-03-17_1 VFPLNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTGDFVKATCEFCGTENLTKEGATTCGYLPQNAVV 186
Jpred
190 200 210 220 230 240 250 260 270
YP_009725298.1 KIYCPACHNSEVGPEHSLAEYHNESGLKTILRKGGRTIAFGGCVFSYVGCHNKCAYWVPRASANIGCNHTGVVGEGSEGLNDNLLEILQKEKV 279
hCoV-19/USA/WA-UW-1741/2020|EPI_ISL_424238|2020-03-21_1 KIYCPACHNSEVGPEHSLAEYHNESGLKTILRKGGRTIAFGGCVFSYVGCHNKCAYWVPRASANIGCNHTGVVGEGSEGLNDNLLEILQKEKV 279
hCoV-19/England/LOND-D4797/2020|EPI_ISL_444153|2020-04-09_1 KIYCPACHNSEVGPEHSLAEYHNESGLKTILRKGGRTIAFGGCVFSYVGCHNKCAYWVPRASANIGCNHTGVVGEGSEGLNDNLLEILQKEKV 279
hCoV-19/USA/CA-SR016/2020|EPI_ISL_430016|2020-03-24_1 KIYCPACHNSEVGPEHSLAEYHNESGLKTILRKGGRTIAFGGCVFSYVGCHNKCAYWVPRASANIGCNHTGVVGEGSEGLN-NLLEILQKEKV 278
hCoV-19/Australia/NSW62/2020|EPI_ISL_417412|2020-03-17_1 KIYCPACHNSEVGPEHSLAEYHNESDLKTILRKGGRTIAFGGCVFSYVGCHNKCAYWVPRASANIGCNHTGVVGEGSEGLNDNLLEILQKEKV 279
Jpred
280 290 300 310 320 330 340 350 360 370
YP_009725298.1 NINIVGDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFKQIVESCGNFKVTKGKAKKGAWNIGEQKSILSPLYAFASEAARVVRSIFSRTL 372
hCoV-19/USA/WA-UW-1741/2020|EPI_ISL_424238|2020-03-21_1 NINIVGDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFKQIVESCGNFKVTKGKAKKGAWNIGEQKSILSPLYAFASEAARVVRSIFSRTL 372
hCoV-19/England/LOND-D4797/2020|EPI_ISL_444153|2020-04-09_1 NINIVGDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFKQIVESCGNFKVTKGKAKKGAWNIGEQKSILSPLYAFASEAARVVRSIFSRTL 372
hCoV-19/USA/CA-SR016/2020|EPI_ISL_430016|2020-03-24_1 NINIVGDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFKQIVESCGNFKVTKGKAKKGAWNIGEQKSILSPLYAFASEAARVVRSIFSRTL 371
hCoV-19/Australia/NSW62/2020|EPI_ISL_417412|2020-03-17_1 NINIVGDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFKQIVESCGNFKVTKGKAKKGAWNIGEQKSILSPLYAFASEAARVVRSIFSRTL 372
Jpred
380 390 400 410 420 430 440 450 460
YP_009725298.1 ETAQNSVRVLQKAAITILDGISQYSLRLIDAMMFTSDLATNNLVVMAYITGGVVQLTSQWLTNIFGTVYEKLKPVLDWLEEKFKEGVEFLRDG 465
hCoV-19/USA/WA-UW-1741/2020|EPI_ISL_424238|2020-03-21_1 ETAQNSVRVLQKAAITILDGISQYSLRLIDAMMFTSDLATNNLVVMAYITGGVVQLTSQWLTNIFGTVYEKLKPVLDWLEEKFKEGVEFLRDG 465
hCoV-19/England/LOND-D4797/2020|EPI_ISL_444153|2020-04-09_1 ETAQNSVRVLQKAAITILDGISQYSLRLIDAMMFTSDLATNNLVVMAYITGGVVQLTSQWLTNIFGTVYEKLKPVLDWLEEKFKEGVEFLRDG 465
hCoV-19/USA/CA-SR016/2020|EPI_ISL_430016|2020-03-24_1 ETAQNSVRVLQKAAITILDGISQYSLRLIDAMMFTSDLATNNLVVMAYITGGVVQLTSQWLTNIFGTVYEKLKPVLDWLEEKFKEGVEFLRDG 464
hCoV-19/Australia/NSW62/2020|EPI_ISL_417412|2020-03-17_1 ETAQNSVRVLQKAAITILDGISQYSLRLIDAMMFTSDLATNNLVVMAYITGGVVQLTSQWLTNIFGTVYEKLKPVLDWLEEKFKEGVEFLRDG 465
Jpred
470 480 490 500 510 520 530 540 550
YP_009725298.1 WEIVKFISTCACEIVGGQIVTCAKEIKESVQTFFKLVNKFLALCADSIIIGGAKLKALNLGETFVTHSKGLYRKCVKSREETGLLMPLKAPKE 558
hCoV-19/USA/WA-UW-1741/2020|EPI_ISL_424238|2020-03-21_1 WEIVKFISTCACEIVGGQIVTCAKEIKESVQTFFKLVNKFLALCADSIIIGGAKLKALNLGETFVTHSKGLYRKCVKSREETGLLMPLKAPKE 558
hCoV-19/England/LOND-D4797/2020|EPI_ISL_444153|2020-04-09_1 WEIVKFISTCACEIVGGQIVTCAKEIKESVQTFFKLVNKFLALCADSIIIGGAKLKALNLGETFVTHSKGLYRKCVKSREETGLLMPLKAPKE 558
hCoV-19/USA/CA-SR016/2020|EPI_ISL_430016|2020-03-24_1 WEIVKFISTCACEIVGGQIVTCAKEIKESVQTFFKLVNKFLALCADSIIIGGAKLKALNLGETFVTHSKGLYRKCVKSREETGLLMPLKAPKE 557
hCoV-19/Australia/NSW62/2020|EPI_ISL_417412|2020-03-17_1 WEIVKFISTCACEIVGGQIVTCAKEIKESVQTFFKLVNKFLALCADSIIIGGAKLKALNLGETFVTHSKGLYRKCVKSREETGLLMPLKAPKE 558
Jpred
560 570 580 590 600 610 620 630
YP_009725298.1 IIFLEGETLPTEVLTEEVVLKTGDLQPLEQPTSEAVEAPLVGTPVCINGLMLLEIKDTEKYCALAPNMMVTNNTFTLKGG 638
hCoV-19/USA/WA-UW-1741/2020|EPI_ISL_424238|2020-03-21_1 IIFLEGETLPTEVLTEEVVLKTGDLQPLEQPTSEAVEAPLVGTPVCINGLMLLEIKDTEKYCALAPNMMVTNNTFTLKGG 638
hCoV-19/England/LOND-D4797/2020|EPI_ISL_444153|2020-04-09_1 VIFLEGETLPTEVLTEEVVLKTGDLQSLEQPTSEAVEAPLVGTPVCINGLMLLEIKDTEKYCALAPNMMVTNNTFTLKGG 638
hCoV-19/USA/CA-SR016/2020|EPI_ISL_430016|2020-03-24_1 IIFLEGETLPTEVLTEEVVLKTGDLQPLEQPTSEAVEAPLVGTPVCINGLMLLEIKDTEKYCALAPNMMVTNNTFTLKGG 637
hCoV-19/Australia/NSW62/2020|EPI_ISL_417412|2020-03-17_1 IIFLEGETLPTEVLTEEVVLKTGDLQPLEQPTSEAVEAPLVGTPVCINGLMLLEIKDTEKYCALAPNMMVTNNTFTLKGG 638
Jpred

Figure 4

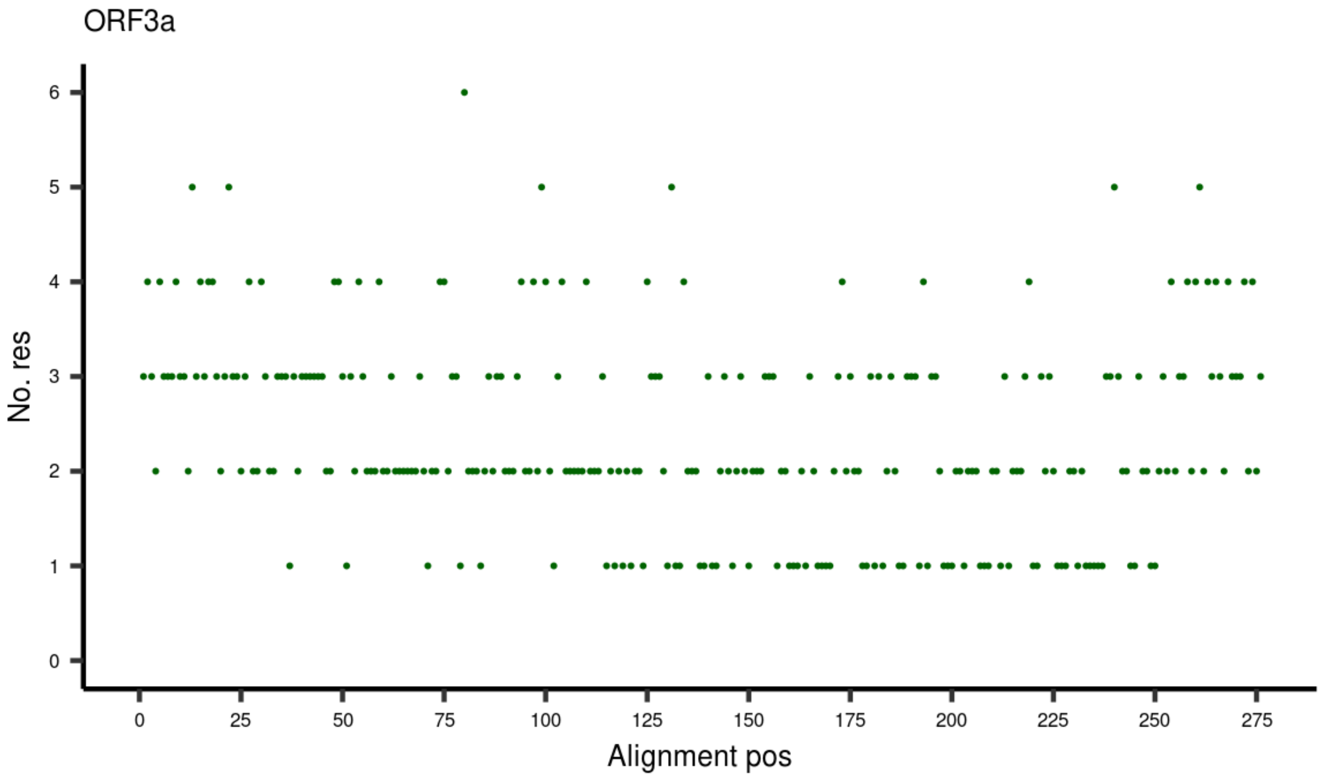

Figure S1

Number of different residues observed in all the variants at each sequence position of ORF3a protein. Number of residue equal to one indicates conservation of the site.

doi:10.20944/preprints202007.0049.v1

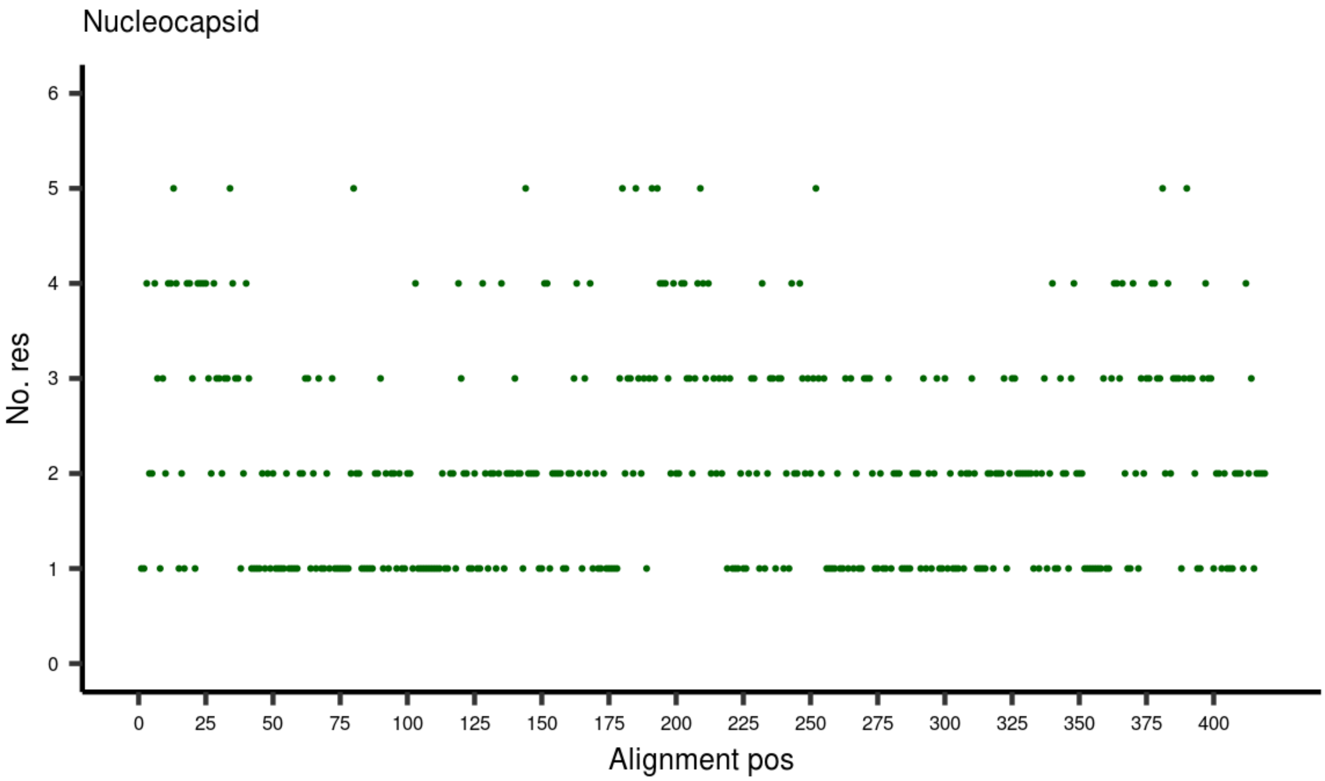

Figure S2

Map of the conserved sites in Nucleocapsid protein. Interpretation of the figure is as described in Figure S1.

doi:10.20944/preprints202007.0049.v1

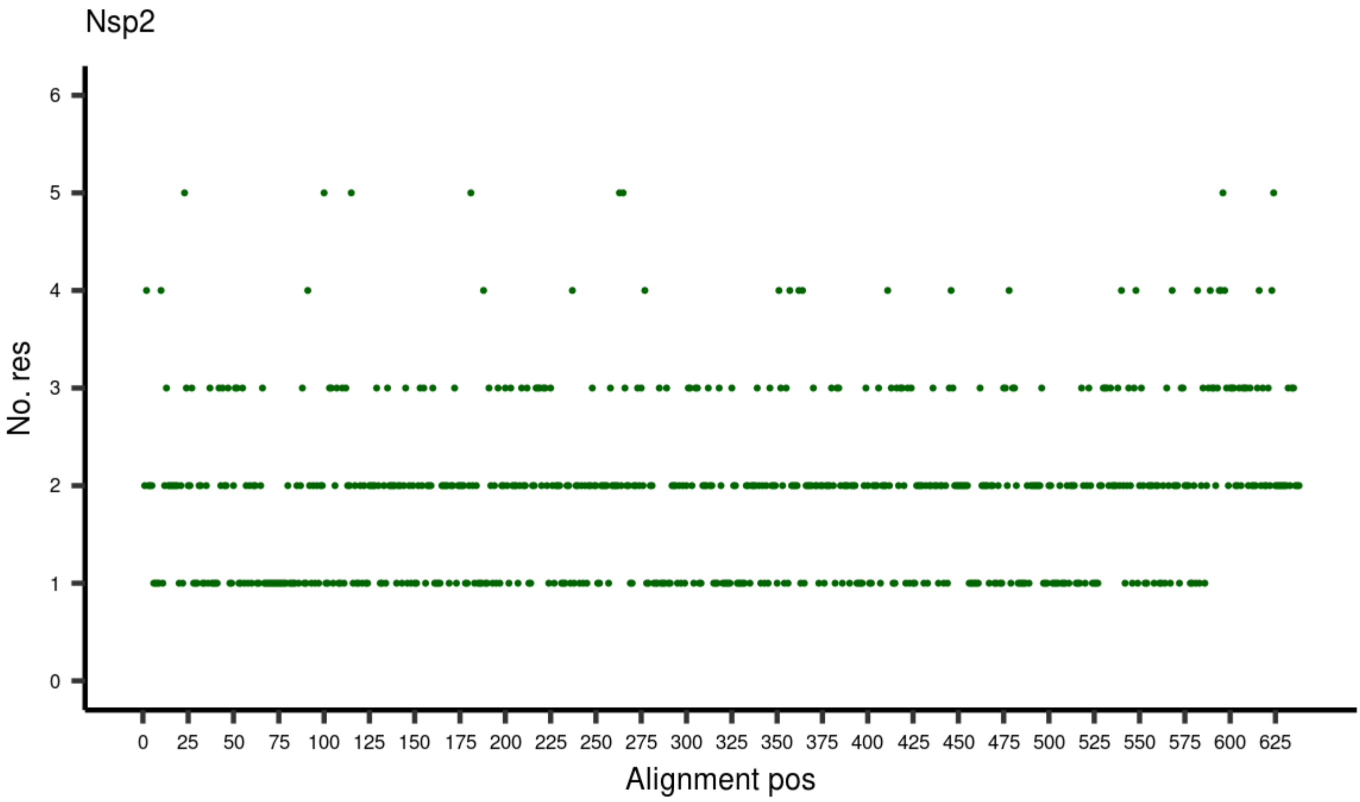

Figure S3

Map of the conserved sites in Nsp2 protein. Interpretation of the figure is as described in Figure S1.